

Sterling K. Berberian

FUNDAMENTALS OF REAL ANALYSIS

$$c = 2^{\aleph_0}$$

$$D \int f = f \text{ a.e.}, \quad (\mathcal{L}^p)' = \mathcal{L}^q$$

$$\int f d(\mu \times \nu) = \iint f d\mu d\nu$$

$$y' = f(x, y), \quad \int (f * g) = (\int f)(\int g)$$

$$(Tf)(x) = \int K(x, y)f(y)dy$$



Springer

Universitext

*Editorial Board
(North America):*

S. Axler
F.W. Gehring
K.A. Ribet

Springer

New York

Berlin

Heidelberg

Barcelona

Hong Kong

London

Milan

Paris

Singapore

Tokyo

Universitext

Editors (North America): S. Axler, F.W. Gehring, and K.A. Ribet

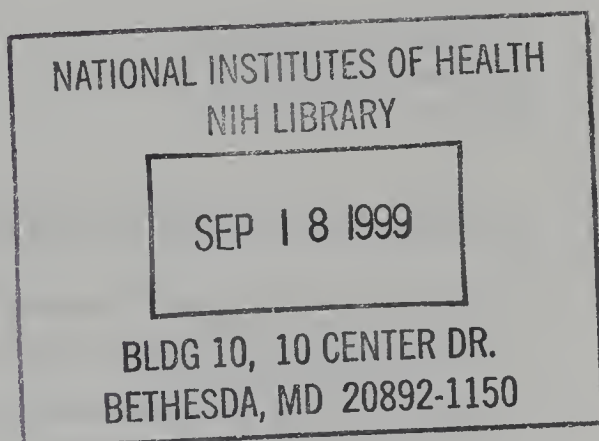
Aksoy/Khamsi: Nonstandard Methods in Fixed Point Theory
Andersson: Topics in Complex Analysis
Aupetit: A Primer on Spectral Theory
Berberian: Fundamentals of Real Analysis
Booss/Bleecker: Topology and Analysis
Borkar: Probability Theory: An Advanced Course
Carleson/Gamelin: Complex Dynamics
Cecil: Lie Sphere Geometry: With Applications to Submanifolds
Chae: Lebesgue Integration (2nd ed.)
Charlap: Bieberbach Groups and Flat Manifolds
Chern: Complex Manifolds Without Potential Theory
Cohn: A Classical Invitation to Algebraic Numbers and Class Fields
Curtis: Abstract Linear Algebra
Curtis: Matrix Groups
DiBenedetto: Degenerate Parabolic Equations
Dimca: Singularities and Topology of Hypersurfaces
Edwards: A Formal Background to Mathematics I a/b
Edwards: A Formal Background to Mathematics II a/b
Foulds: Graph Theory Applications
Friedman: Algebraic Surfaces and Holomorphic Vector Bundles
Fuhrmann: A Polynomial Approach to Linear Algebra
Gardiner: A First Course in Group Theory
Gårding/Tambour: Algebra for Computer Science
Goldblatt: Orthogonality and Spacetime Geometry
Gustafson/Rao: Numerical Range: The Field of Values of Linear Operators and Matrices
Hahn: Quadratic Algebras, Clifford Algebras, and Arithmetic Witt Groups
Holmgren: A First Course in Discrete Dynamical Systems
Howe/Tan: Non-Abelian Harmonic Analysis: Applications of $SL(2, \mathbb{R})$
Howes: Modern Analysis and Topology
Humi/Miller: Second Course in Ordinary Differential Equations
Hurwitz/Kritikos: Lectures on Number Theory
Jennings: Modern Geometry with Applications
Jones/Morris/Pearson: Abstract Algebra and Famous Impossibilities
Kannan/Krueger: Advanced Analysis
Kelly/Matthews: The Non-Euclidean Hyperbolic Plane
Kostrikin: Introduction to Algebra
Luecking/Rubel: Complex Analysis: A Functional Analysis Approach
MacLane/Moerdijk: Sheaves in Geometry and Logic
Marcus: Number Fields
McCarthy: Introduction to Arithmetical Functions
Meyer: Essential Mathematics for Applied Fields
Mines/Richman/Ruitenburg: A Course in Constructive Algebra
Moise: Introductory Problems Course in Analysis and Topology
Morris: Introduction to Game Theory
Polster: A Geometrical Picture Book
Porter/Woods: Extensions and Absolutes of Hausdorff Spaces
Ramsay/Richtmyer: Introduction to Hyperbolic Geometry
Reisel: Elementary Theory of Metric Spaces
Rickart: Natural Function Algebras

(continued after index)

Sterling K. Berberian

Fundamentals of Real Analysis

With 31 Figures



Springer

QA
300
B4574
1999
Sterling K. Berberian
Department of Mathematics
University of Texas at Austin
Austin, TX 78712-1082
USA

*Editorial Board
(North America):*

S. Axler
Mathematics Department
San Francisco State University
San Francisco, CA 94132
USA

F.W. Gehring
Mathematics Department
East Hall
University of Michigan
Ann Arbor, MI 48109-1109
USA

K.A. Ribet
Department of Mathematics
University of California at Berkeley
Berkeley, CA 94720-3840
USA

Mathematics Subject Classification (1991): 26, 28, 46, 54, 04

Library of Congress Cataloging-in-Publication Data
Berberian, Sterling K., 1926–
Fundamentals of real analysis / Sterling K. Berberian.
p. cm. —(Universitext)
Includes bibliographical references and indexes.
ISBN 0-387-98480-1 (pbk. : alk. paper)
1. Mathematical analysis. I. Title.
QA300.B4574 1998
515—dc21 98-13045

Printed on acid-free paper.

© 1999 Springer-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Anthony K. Guardiola; manufacturing supervised by Jeffrey Taub. Photocomposed copy using the author's \TeX files prepared by TechBooks, Fairfax, VA. Printed and bound by Edwards Brothers, Inc., Ann Arbor, MI.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-98480-1 Springer-Verlag New York Berlin Heidelberg SPIN 10668240

*To the memory of James Ellis Powell,
late Professor Emeritus, Michigan State University*

Preface

This book is a record of a course on functions of a real variable, addressed to first-year graduate students in mathematics, offered in the academic year 1985–86 at the University of Texas at Austin. It consists essentially of the day-by-day lecture notes that I prepared for the course, padded up with the exercises that I seemed never to have the time to prepare in advance; the structure and contents of the course are preserved faithfully, with minor cosmetic changes here and there.

Two facts are worth noting: (1) the lecture notes were prepared (if not always delivered) with exceptional care, as my son was enrolled in the class and I confess that I was trying especially hard to put my best foot forward; (2) the text does not reflect the fact that I wasted a certain amount of time doing Lebesgue’s “Fundamental theorem of calculus” at the end of the first semester, ‘discovered’ E.J. McShane’s lovely exposition during the semester break, and was so struck by the superiority of his exposition that I did the topic all over again at the beginning of the second semester. It is only the ‘second pass’ that is recorded here (in Chapter 5); the time saved by doing it right in the first place should be ample for including the very few topics I added that were not covered in the actual course (notably, the Riesz representation theorem, included here as Theorem 6.7.11—the 11th item in §7 of Chapter 6).

The choice of topics and the order in which they are taken up was guided by the following principles:

(1) The most important things should come first (it is a little intellectually arrogant to make such judgments, but that’s what a teacher is paid to do—and the student need not, and sometimes should not, agree). When planning the course, at each topic I kept in mind the question: “If the student is obliged to drop out tomorrow—or who takes only the first semester, as is frequently the case—will he or she have been exposed to the topics that are most likely to be crucial in his or her mathematical development?”

(2) Every subject becomes fatiguing after a while, and when fatigue sets in, learning converges rapidly to zero. For example, the course syllabus called for a full-dress treatment of measure and integration, but consuming it all in one gulp leads to indigestion (I ask forgiveness of all the students on whom I inflicted one-semester or even one-year courses in Measure and

Integration; we got some good out of it and I amassed enough material for a book on the subject, but it was not the best use of our time). Therefore, the theme of measure theory must be broken up into digestible units and alternated with other themes for the sake of variety. The same is true of topology and function spaces: a generous portion, but not all in one gulp.

(3) The house being built, to be sturdy and serviceable, must have a foundation: the first part of the course must come to grips with the real numbers (they have to be constructed rigorously from the rationals), the axioms of set theory (just visiting!) and the concepts of cardinality and ordinality (indispensable tools in grappling with infinity, one of the mathematician's principal occupations); for an eloquent essay on the importance of taking up such matters, I refer the reader to the Preface of Irving Kaplansky's *Set theory and metric spaces* [2nd edn., Chelsea, New York, 1977].

A certain amount of inefficiency is introduced in the passage from concrete to abstract (measure spaces), special to general (metric and topological spaces), finite to infinite (product measure, signed measures), real to complex (function spaces), and so on. This seemed not burdensome in the classroom, where a few words often sufficed to reset the stage for the reappearance of a subject, but in print it is necessary to revisit a considerable amount of notation and definitions, especially when related discussions are widely separated in time (pages). The benefits of recurrent themes (motivation, boredom avoidance) seemed worth the inefficiency in class; I hope the reader will find that they also make the book easier to read.

Can the topics taken up be treated more effectively? Assuredly. Could I have chosen more important topics to take up? At the time, I thought not, and, a decade later, I feel sufficiently comfortable with the choices to warrant putting the lecture notes into a more presentable form; the ultimate verdict, as always, is the reader's.

Austin, Texas
September 1996

Sterling K. Berberian

Contents

Preface	vii
-------------------	-----

CHAPTER 1

Foundations	1
§1.1. Logic, set notations	2
§1.2. Relations	5
§1.3. Functions (mappings)	9
§1.4. Product sets, axiom of choice	12
§1.5. Inverse functions	15
§1.6. Equivalence relations, partitions, quotient sets	17
§1.7. Order relations	20
§1.8. Real numbers	26
§1.9. Finite and infinite sets	34
§1.10. Countable and uncountable sets	38
§1.11. Zorn's lemma, the well-ordering theorem	41
§1.12. Cardinality	46
§1.13. Cardinal arithmetic, the continuum hypothesis	52
§1.14. Ordinality	59
§1.15. Extended real numbers	73
§1.16. limsup, liminf, convergence in $\overline{\mathbb{R}}$	79

CHAPTER 2

Lebesgue Measure	86
§2.1. Lebesgue outer measure on \mathbb{R}	86
§2.2. Measurable sets	92
§2.3. Cantor set: an uncountable set of measure zero	98
§2.4. Borel sets, regularity	101
§2.5. A nonmeasurable set	109
§2.6. Abstract measure spaces	112

CHAPTER 3

Topology	115
§3.1. Metric spaces: examples	116
§3.2. Convergence, closed sets and open sets in metric spaces	123

§3.3.	Topological spaces	130
§3.4.	Continuity	138
§3.5.	Limit of a function	141

CHAPTER 4

Lebesgue Integral	148
§4.1. Measurable functions	149
§4.2. a.e.	156
§4.3. Integrable simple functions	160
§4.4. Integrable functions	164
§4.5. Monotone convergence theorem, Fatou's lemma	173
§4.6. Monotone classes	178
§4.7. Indefinite integrals	184
§4.8. Finite signed measures	189

CHAPTER 5

Differentiation	199
§5.1. Bounded variation, absolute continuity	201
§5.2. Lebesgue's representation of AC functions	213
§5.3. limsup, liminf of functions; Dini derivates	215
§5.4. Criteria for monotonicity	222
§5.5. Semicontinuity	229
§5.6. Semicontinuous approximations of integrable functions	239
§5.7. F. Riesz's "Rising sun lemma"	242
§5.8. Growth estimates of a continuous increasing function	246
§5.9. Indefinite integrals are a.e. primitives	248
§5.10. Lebesgue's "Fundamental theorem of calculus"	252
§5.11. Measurability of derivates of a monotone function	253
§5.12. Lebesgue decomposition of a function of bounded variation	257
§5.13. Lebesgue's criterion for Riemann-integrability	265

CHAPTER 6

Function Spaces	273
§6.1. Compact metric spaces	273
§6.2. Uniform convergence, iterated limits theorem	285
§6.3. Complete metric spaces	299
§6.4. L^1	311
§6.5. Real and complex measures	319
§6.6. L^∞	323
§6.7. L^p ($1 < p < \infty$)	332
§6.8. $\mathcal{C}(X)$	345
§6.9. Stone–Weierstrass approximation theorem	353

CHAPTER 7

Product Measure 364

§7.1. Extension of measures 365

§7.2. Product measures 371

§7.3. Iterated integrals, Fubini–Tonelli theorem for
finite measures 382

§7.4. Fubini–Tonelli theorem for σ -finite measures 388

CHAPTER 8

The Differential Equation $y' = f(x, y)$ 398

§8.1. Equicontinuity, Ascoli’s theorem 398

§8.2. Picard’s existence theorem for $y' = f(x, y)$ 408

§8.3. Peano’s existence theorem for $y' = f(x, y)$ 416

CHAPTER 9

Topics in Measure and Integration 422

§9.1. Jordan–Hahn decomposition of a signed measure 422

§9.2. Radon–Nikodym theorem 432

§9.3. Lebesgue decomposition of measures 444

§9.4. Convolution in $L^1(\mathbb{R})$ 451

§9.5. Integral operators (with continuous kernel function) 460

Bibliography 469

Index of Notations 471

Index 473

CHAPTER 1

Foundations

- §1.1. Logic, set notations
- §1.2. Relations
- §1.3. Functions (mappings)
- §1.4. Product sets, Axiom of Choice
- §1.5. Inverse functions
- §1.6. Equivalence relations, Partitions, Quotient Sets
- §1.7. Order relations
- §1.8. Real numbers
- §1.9. Finite and infinite sets
- §1.10. Countable and uncountable sets
- §1.11. Zorn's Lemma, the Well-Ordering theorem
- §1.12. Cardinality
- §1.13. Cardinal arithmetic
- §1.14. Ordinality
- §1.15. Extended real numbers
- §1.16. Convergence in $\overline{\mathbb{R}}$

The reader will already have a working familiarity with the concepts of set and function. Apart from a review of basic concepts and notations, the chapter is mostly about coping with infinity (or exploiting it—it depends on one's point of view). Our viewpoint (generally called 'naive') is that infinite sets exist and they are not to be feared. Some of the axioms of set theory (the Axiom of Choice, the Continuum Hypothesis) are more controversial than the others; whether or not one admits them is a matter of professional lifestyle.¹ In this text, the axiom of choice and its logically equivalent forms (§1.11) are admitted and are invoked whenever convenient. The continuum hypothesis is only mentioned briefly (§1.13); although it is not used anywhere in the text, it is instructive to understand the terms needed to state it.

¹ Assuming the usual (Zermelo-Fraenkel) axioms of set theory are consistent [cf. I. Kaplansky, *Set theory and metric spaces*, Chapter 3, 2nd edn., Chelsea, New York, 1977]. If ZF goes down the tube, we all go down with it.

1.1. Logic, Set Notations¹

1.1.1. A 'short list' of useful symbols:

<i>Notation</i>	<i>Read</i>
$x \in A$	x is an element of the set A
\forall	for all, for every
\exists	there exists (at least one)
$\exists!$	there exists a unique (one and only one)
\ni	such that (having the following properties)
$\&, \wedge$	and
\vee	or (non-exclusive)
\Rightarrow	implies
\Leftrightarrow	if and only if (a fusion of \Rightarrow and \Leftarrow)
\sim	negation (of a proposition)

1.1.2. A *proposition* is a statement that is either true or false (but not both). If P is a proposition, its *negation* $\sim P$ is the proposition that is false when P is true, true when P is false. For example, in ordinary arithmetic, if P is the (false) proposition $\ll 3 \leq 2 \gg$ then $\sim P$ is the (true) proposition $\ll 2 < 3 \gg$; more generally, if P is $\ll x \leq y \gg$ then $\sim P$ is $\ll y < x \gg$.

1.1.3. Before explaining the usage of the other symbols, it helps to have a repertory of specific sets:

<i>Symbol</i>	<i>Meaning</i>
\mathbb{P}	the set $\{1, 2, 3, \dots\}$ of all <i>positive integers</i>
\mathbb{N}	the set $\{0, 1, 2, 3, \dots\}$ of all <i>nonnegative integers</i>
\mathbb{Z}	the set $\{0, \pm 1, \pm 2, \pm 3, \dots\}$ of all <i>integers</i>
\mathbb{Q}	the set of all <i>rational numbers</i> m/n ($m, n \in \mathbb{Z}; n \neq 0$)
\mathbb{R}	the set of all <i>real numbers</i>
\mathbb{C}	the set of all <i>complex numbers</i> $z = x + iy$ ($x, y \in \mathbb{R}, i^2 = -1$)

The construction of \mathbb{R} from the rational field \mathbb{Q} is sketched in §1.8; the construction of \mathbb{C} from \mathbb{R} is elementary algebra. The notations \mathbb{R} and \mathbb{C} are standard; \mathbb{N} , \mathbb{Z} and \mathbb{Q} are 'fairly standard' (i.e., widely used); \mathbb{P} is improvised (no consensus!).

1.1.4. If P and Q are propositions, then $P \Rightarrow Q$ means that if P is true then Q is also true. For example, the statement

$$x \in \mathbb{Z} \Rightarrow x \in \mathbb{Q}$$

¹I suggest first glancing at the tables in this section; if everything looks familiar, the section can be omitted.

says that every integer is a rational number; it is true. The *converse* statement (with implication pointing in the reverse direction)

$$x \in \mathbb{Z} \Leftarrow x \in \mathbb{Q}$$

happens to be false, but it is a legitimate statement. (The rules of ordinary language do not abolish lies.) When we demonstrate that $P \Rightarrow Q$, we say that we have proved a *theorem*, with *hypothesis* P and *conclusion* Q . Sometimes (often!) a theorem $P \Rightarrow Q$ is proved by showing that $\sim Q \Rightarrow \sim P$ (the *contrapositive* form of $P \Rightarrow Q$).

1.1.5. If propositions P and Q imply each other, they are said to be logically *equivalent*, written $P \Leftrightarrow Q$ (or $P \equiv Q$); thus

$$(P \Leftrightarrow Q) \equiv (P \Rightarrow Q) \& (Q \Rightarrow P).$$

For example, in ordinary arithmetic,

$$x = y \Leftrightarrow (x \leq y) \& (y \leq x).$$

The basis of proofs in contrapositive form is the equivalence

$$(P \Rightarrow Q) \equiv (\sim Q \Rightarrow \sim P).$$

An equivalence of some depth: for a real number x ,

$$x \geq 0 \Leftrightarrow (\exists y \in \mathbb{R} \ni x = y^2)$$

(translate it from ‘symbolese’ into ordinary language!).

1.1.6. The symbol \forall is sometimes used literally, sometimes as a ‘stage-setter’ (or ‘quantifier’) indicating the set in which a statement is formulated. Consider, for example, the statements

$$x^2 \geq 0 \quad (\forall x \in \mathbb{R})$$

$$(\forall x \in \{1, 2, 3, 4\}) x = 1 \Leftrightarrow x^2 < 4.$$

In the first example, the condition on x ($x^2 \geq 0$) is true for every x in the set \mathbb{R} , thus the statement simply says that x^2 is nonnegative for *every* real number x ; here, \forall is used in its ordinary, literal sense. In the second example, the condition on x (the assertion of an equivalence \Leftrightarrow) is also true for every x in the set $\{1, 2, 3, 4\}$ (albeit in a vacuous way for $x = 2, 3$ or 4), though its constituent pieces ($x = 1$, $x^2 < 4$) are not.

1.1.7. The mathematical ‘or’ is used ‘permissively’ rather than ‘exclusively’; thus, the statement

$$(x \in A) \vee (x \in B)$$

does not exclude the possibility that both $x \in A$ and $x \in B$.

1.1.8. If A and B are sets such that $x \in A \Rightarrow x \in B$, then A is called a *subset* of B , written $A \subset B$ (alternatively, B is a *superset*

of A , written $B \supset A$). For example, the set E of even integers is a subset of \mathbb{Z} ; it can be specified as the set of all integers n such that $n = 2k$ for some integer k , a recipe conveniently expressed by

$$E = \{n \in \mathbb{Z} : n = 2k \text{ for some } k \in \mathbb{Z}\}$$

(the colon is read as “such that”). More generally, if X is a set and if, for each $x \in X$, $P(x)$ is a proposition involving x , then

$$\{x \in X : P(x)\}$$

denotes the set of all elements x of X for which $P(x)$ is true; this can be shortened to $\{x : P(x)\}$ when there is no doubt as to the ‘universal set’ X from which the elements x are drawn. For example,

$$\{x \in \mathbb{Z} : -2 \leq x < 4\} = \{-2, -1, 0, 1, 2, 3\}$$

(a set with six elements), whereas

$$\{x \in \mathbb{R} : -2 \leq x < 4\} = [-2, 4)$$

(a semi-closed interval); unless a universal set (such as \mathbb{Z} or \mathbb{R}) is specified, the notation $\{x : -2 \leq x < 4\}$ is ambiguous. An expression such as $\{x : x \in A \ \& \ x \in B\}$ is unambiguous, since it can be rewritten as $\{x \in A : x \in B\}$.

1.1.9. Fix a universal set X and let A, B, C, \dots be subsets of X . The following table lists the most basic set-theoretic notations (others follow in later sections):

<i>Symbol</i>	<i>Meaning</i>
$x \notin A$	$\sim (x \in A)$ (that is, x <i>not</i> an element of A)
$A \subset B$	$x \in A \Rightarrow x \in B$
$A \not\subset B$	$\sim (A \subset B)$ (that is, $\exists x \in A \ni x \notin B$)
$A = B$	$x \in A \Leftrightarrow x \in B$ (that is, $A \subset B$ and $B \subset A$)
$A \neq B$	$\sim (A = B)$ (that is, either $A \not\subset B$ or $B \not\subset A$)
$A \subsetneq B$	$A \subset B \ \& \ A \neq B$ (that is, $A \subset B$ and $B \not\subset A$; A is then said to be a <i>proper</i> subset of B , and B is said to contain A <i>properly</i>)
$B \supset A$	$A \subset B$
$A \cap B$	$\{x : x \in A \ \& \ x \in B\}$ (the <i>intersection</i> of A and B)
$A \cup B$	$\{x : x \in A \text{ or } x \in B\}$ (the <i>union</i> of A and B)
$\complement_X A$	$\{x \in X : x \notin A\}$ (the <i>complement</i> of A in X)
$A - B$	$\{x : x \in A \ \& \ x \notin B\}$ (the <i>difference</i> ‘ A minus B ’, or the ‘relative complement’ of B in A)

The complement of A in X is also written $\complement A$ or A' ; thus $A - B = A \cap \complement B = A \cap B'$, $(A')' = A$, $A \cup A' = X$ and $A \cap A' = \emptyset$ (the *empty*

set). Some other useful formulas are listed in the exercises for convenient reference.

Finally, a caution about the use of the words ‘all’ and ‘every’:

1.1.10. (*Russell’s paradox*)² The statement “There exists a set of which every set is a member” is nonsense. For, if U were such a set, then its subset

$$A = \{x \in U : x \notin x\}$$

would be a member of U , leading inexorably to a contradiction: either $A \in A$ (in which case $A \notin A$ by the definition of A), or $A \notin A$ (in which case $A \in A$ by the definition of A).

Moral. The words ‘all’ and ‘every’ are very big (too big); to play it safe, qualify by operating within a known set. The following usage of ‘all’ is prudent: ‘The set of all one-element subsets of \mathbb{P} ’; the sets A in question are qualified by the condition $A \subset \mathbb{P}$. The expression ‘The set of all one-element sets’ is asking for trouble. {Trouble: Let E be ‘the set of all one-element sets’, then consider the set F of all sets A that contain an element of E (in other words, $A \neq \emptyset$); we are now face to face with $F \cup \{\emptyset\}$, the dreaded ‘set of all sets’.

Exercises

1. Let X be a set, A, B, C subsets of X , A' the complement of A .

(i) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(i') $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(ii) $A \subset B \Leftrightarrow A' \supset B'$

(iii) $(A \cup B)' = A' \cap B'$

(iii') $(A \cap B)' = A' \cup B'$

(iv) $A \subset B \Leftrightarrow A = A \cap B$

(iv') $A \subset B \Leftrightarrow B = A \cup B$

2. The description of a “proposition” in 1.1.2 can be expressed as follows: For every proposition P , $P \vee (\sim P)$ is true (*law of the excluded middle*) and $P \wedge (\sim P)$ is false (*law of contradiction*).

1.2. Relations

1.2.1. *Definition.* If X and Y are sets, the **cartesian product** of X and Y (in that order), denoted $X \times Y$, is the set of all ordered pairs

² Bertrand Russell (1872-1970).

(x, y) with $x \in X$ and $y \in Y$:

$$X \times Y = \{(x, y) : x \in X \text{ \& } y \in Y\},$$

with the understanding that

$$(x, y) = (x', y') \Leftrightarrow x = x' \text{ \& } y = y'.$$

One calls x and y the first and second *coordinates* of (x, y) (cf. Figure 1).

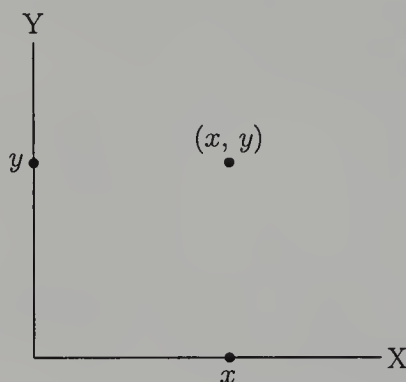


Figure 1

1.2.2. *Definition.* A **relation** from X to Y (in that order) is a subset R of $X \times Y$:

$$R \subset X \times Y$$

(cf. Figure 2). If $(x, y) \in R$ we write xRy (read “ x is related by R to y ”), and if $(x, y) \notin R$ we write $xR'y$ (an appropriate notation, since (x, y) belongs to the complement R' of R). If $X = Y$ we say that R is a relation *in* X .

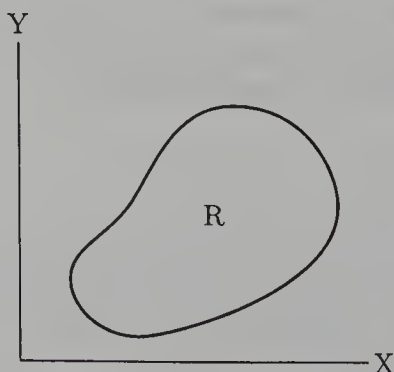


Figure 2

1.2.3. *Example.* Let $X = \{1, 2, 3, 4\}$ and let R be the usual relation “ $<$ ” in X ; as a set of ordered pairs,

$$R = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}.$$

1.2.4. *Example.* Let $X = \{1, 2, 3, 4\}$ and suppose that xRy means that $x|y$ (x is a divisor of y). Then

$$R = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 4), (3, 3), (4, 4)\}.$$

1.2.5. *Example.* If R is a relation in X and A is a subset of X , then $R \cap (A \times A)$ is a relation in A , said to be **induced** in A by R .

1.2.6. *Definition.* Let R be a relation from X to Y (1.2.2). For each subset A of X , we write

$$R(A) = \{y \in Y : xRy \text{ for some } x \in A\}$$

and call it the (direct) **image** of A under R ; for each subset B of Y , we write

$$R^{-1}(B) = \{x \in X : xRy \text{ for some } y \in B\}$$

and call it the **inverse image** of B under R (cf. Figure 3).

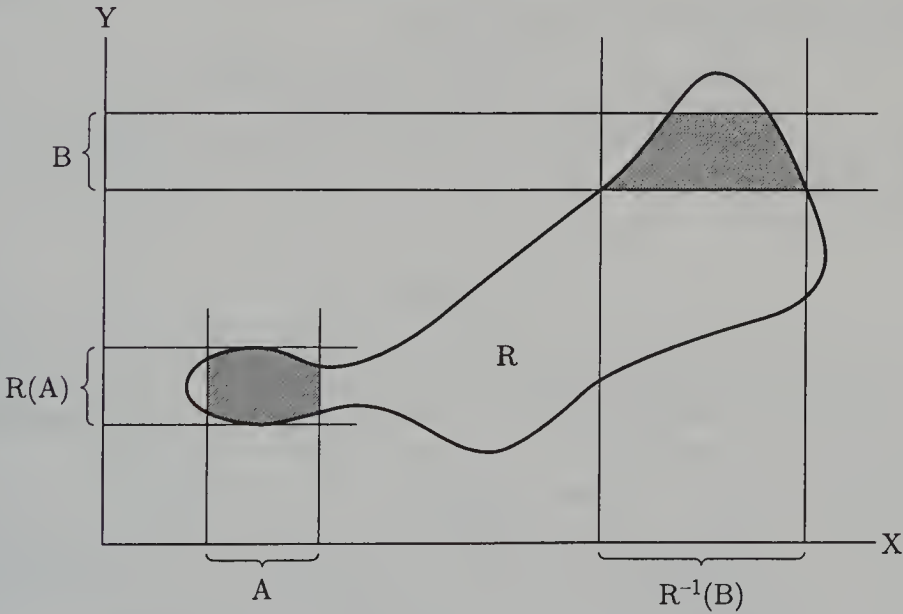


Figure 3

1.2.7. *Remarks.* With notations as in 1.2.6, one can think of R^{-1} as a relation from Y to X , where

$$yR^{-1}x \Leftrightarrow xRy,$$

that is,

$$R^{-1} = \{(y, x) : (x, y) \in R\};$$

R^{-1} is called the **reverse** (or *opposite*) of the relation R . For a subset B of Y , its inverse image under R coincides with its direct image under R^{-1} . Also, $(R^{-1})^{-1} = R$, thus $R(A)$ coincides with the inverse image of A under R^{-1} .

1.2.8. *Example.* If R is the usual relation “ $<$ ” in the set of integers, then R^{-1} is the relation “ $>$ ”, whereas the relation R' of 1.2.2 is the relation “ \geq ”.

1.2.9. *Remark.* With notations as in 1.2.6, if $a \in X$ and $A = \{a\}$ is the set whose only element is a (such sets are called *singletons*) then the image

$$R(\{a\}) = \{y \in Y : (a, y) \in R\}$$

is the set of all second coordinates of the points of the intersection $(\{a\} \times Y) \cap R$ (called the *a-slice* of R).

1.2.10. *Definition.* If R is a relation from X to Y , the set

$$\begin{aligned} R^{-1}(Y) &= \{x \in X : xRy \text{ for some } y \in Y\} \\ &= \{x \in X : R(\{x\}) \neq \emptyset\} \end{aligned}$$

is called the **domain** of R , and the set

$$\begin{aligned} R(X) &= \{y \in Y : xRy \text{ for some } x \in X\} \\ &= \{y \in Y : R^{-1}(\{y\}) \neq \emptyset\} \end{aligned}$$

is called the **range** of R .

Exercises

1. Let R be a relation from X to Y . Prove:

(i) $R(A \cup B) = R(A) \cup R(B)$ and $R(A \cap B) \subset R(A) \cap R(B)$ for every pair of subsets A, B of X .

(ii) The inclusion in (i) may be proper.

2. Let R be a relation from X to Y and let S be a relation from Y to Z . The relation $R \circ S$, called the *composite* of R and S (in that order), is the relation from X to Z defined by the formula

$$R \circ S = \{(x, z) \in X \times Z : \exists y \in Y \ni xRy \text{ and } yRz\}.$$

(i) In addition to the foregoing notations, suppose that T is a relation from Z to W . Prove that $(R \circ S) \circ T = R \circ (S \circ T)$.

(ii) Suppose $Y = X$, so that R is a relation from X to X . Let $\Delta = \{(x, x) : x \in X\}$ be the ‘diagonal’ of $X \times X$ (also called the *identity relation* in X). Note that the condition $\Delta \subset R$ signifies that xRx for all $x \in X$. Express in terms of the xRy notation the meanings of each of

the conditions $R^{-1} = R$ and $R \circ R \subset R$. (A relation satisfying all three of these conditions is called an *equivalence relation*.)

1.3. Functions (Mappings)

1.3.1. *Definition.* Let X and Y be *nonempty* sets. A **function** from X to Y is a relation F from X to Y such that (i) the domain of F is all of X , and (ii) if xFy and xFz then $y = z$.¹

Condition (i) says that, for every $x \in X$, xFy for *at least* one $y \in Y$, whereas (ii) says that xFy for *at most* one $y \in Y$; together, they say that xFy for *exactly* one $y \in Y$, that is,

$$(\forall x \in X) \exists! y \in Y \ni xFy,$$

in other words, for each $x \in X$, $F(\{x\})$ is a singleton. The unique element y such that $F(\{x\}) = \{y\}$ is called the *image* of x under F (or the *value* of F at x) and is denoted $F(x)$. As a set of ordered pairs,

$$F = \{(x, y) \in X \times Y : y = F(x)\}.$$

We write

$$F : X \rightarrow Y$$

to indicate that F is a function from X to Y ; the domain X of F is also called the **initial set** of F , and Y the **final set** of F . Functions $F : X \rightarrow Y$ and $G : X' \rightarrow Y'$ are regarded as *equal* if $X = X'$, $Y = Y'$ and $F(x) = G(x)$ for all $x \in X$. If $F : X \rightarrow Y$, the subset

$$\{(x, F(x)) : x \in X\}$$

of $X \times Y$ is called the **graph** of F ; if $Y' \supset Y$ properly and $F' : X \rightarrow Y'$ is defined by $F'(x) = F(x)$ ($\forall x \in X$), then F and F' have the same graph but technically they are regarded as distinct functions.² The symbol $X \rightarrow Y$ indicates a function from X to Y without specifying a letter F, G , etc. for the function. In the same vein, $x \mapsto F(x)$ indicates the effect of a function F on an element x of its domain, verbalized as ' F sends x to $F(x)$ ', also expressed by the notation $F : x \mapsto F(x)$. For example, the sine function $x \mapsto \sin x$ is a function $\mathbb{R} \rightarrow [-1, 1]$, and $x \mapsto x^2 + 5$ defines a quadratic polynomial function $\mathbb{R} \rightarrow \mathbb{R}$. Some synonyms for "function": **mapping, transformation, operator**.

¹The convention that X and Y are nonempty is probably a minority position. The rationale for adhering to it: a function is generally expected to *do* something. If the domain of a proposed function is not known at the outset to be nonempty, it seems prudent to not let it slip into the discussion, where it may, inadvertently, do mischief.

²This apparent fussiness is important for 'diagrams' (coherent arrays of functions—for example, as in Figures 4 and 5 of the present section).

1.3.2. *Definition.* A function $f : X \rightarrow Y$ is said to be

- (i) **injective** (or *one-one*) if it maps distinct elements of X to distinct elements of Y , that is, $x \neq x' \Rightarrow f(x) \neq f(x')$ (stated contrapositively, $f(x) = f(x') \Rightarrow x = x'$);
- (ii) **surjective** (or *onto*) if its range is all of Y , that is, $f(X) = Y$;
- (iii) **bijective** (or to be a *one-one correspondence*) if it is both injective and surjective.

An injective function is called an **injection**; the terms **surjection** and **bijection** are defined similarly. If there exists a bijection $X \rightarrow Y$, we say that X is **bijective with** Y .

1.3.3. *Examples.* The mapping $[-\pi/2, \pi/2] \rightarrow \mathbb{R}$ defined by $x \mapsto \sin x$ is injective but not surjective; the mapping $\mathbb{R} \rightarrow [-1, 1]$ defined by $x \mapsto \sin x$ is surjective but not injective; the mapping $[-\pi/2, \pi/2] \rightarrow [-1, 1]$ defined by $x \mapsto \sin x$ is bijective; and the mapping $\mathbb{R} \rightarrow \mathbb{R}$ defined by $x \mapsto \sin x$ is none of the above.

1.3.4. *Examples.* For every nonempty set X , the mapping $X \rightarrow X$ defined by $x \mapsto x$ is called the **identity mapping** of X and is denoted id_X ; thus, $\text{id}_X : X \rightarrow X$, $\text{id}_X(x) = x$ ($\forall x \in X$). More generally, if A is a nonempty subset of X , the mapping $A \rightarrow X$ defined by $x \mapsto x$ ($x \in A$) is called the **insertion mapping** of A into X , denoted $i_A : A \rightarrow X$. In particular, $i_X = \text{id}_X$. The function $\varphi_A : X \rightarrow \{0, 1\}$ defined by

$$\varphi_A(x) = \begin{cases} 1 & \text{for } x \in A \\ 0 & \text{for } x \in \complement A \end{cases}$$

is called the **characteristic function** of A ; φ_A is often regarded as having values in the field \mathbb{R} of real numbers (cf. Chapter 4), but there is no law against letting 0, 1 be elements of any field, for example the field of two elements.

1.3.5. *Definition.* Given a pair of functions $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ (g picks up where f leaves off), the function $X \rightarrow Z$ defined by $x \mapsto g(f(x))$ is called the **composite** of g and f and is denoted $g \circ f$ (see Figure 4),

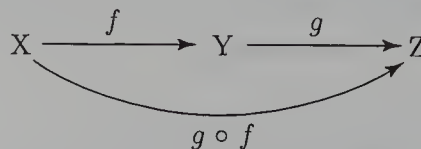


Figure 4

thus $g \circ f : X \rightarrow Z$ and $(g \circ f)(x) = g(f(x))$ ($\forall x \in X$).

1.3.6. *Example.* If $f : X \rightarrow Y$ and A is a nonempty subset of X , the correspondence $x \mapsto f(x)$ ($x \in A$) defines a function $A \rightarrow Y$, called the **restriction** of f to A and denoted $f|A$. Thus $f|A : A \rightarrow Y$ and $(f|A)(x) = f(x)$ ($\forall x \in A$).

Note that the insertion mapping $i_A : A \rightarrow X$ is precisely the restriction mapping $\text{id}_X|A$, while $f|A$ is the composite $f \circ i_A$ (Figure 5).

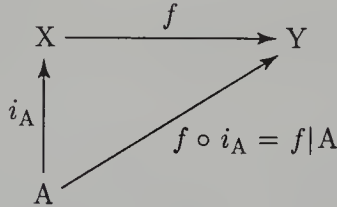


Figure 5

1.3.7. *Theorem.* If $f : X \rightarrow Y$, $g : Y \rightarrow X$ and $g \circ f = \text{id}_X$, then
(i) f is injective, and (ii) g is surjective.

Proof. (i) If $f(x) = f(x')$ then $g(f(x)) = g(f(x'))$, whence $x = x'$.

(ii) For all $x \in X$, $x = g(f(x)) \in g(Y)$, thus $X = g(Y)$. \diamond

A slight revision of the function notation provides a useful generalization of the familiar concept of sequence:

1.3.8. *Definition.* Let I and X be nonempty sets. A function $f : I \rightarrow X$ is also called a **family** of elements of X , **indexed** by I ; with this terminology, one often writes $x_i = f(i)$ for $i \in I$ and one speaks of $(x_i)_{i \in I}$ as being the family and I as the **index set**. Thus, $(x_i)_{i \in I}$ is a substitute notation for a function $i \mapsto x_i$ ($i \in I$).

1.3.9. *Examples.* A family of elements of X indexed by \mathbb{P} or \mathbb{N} is called a *sequence* of elements of X ; a family indexed by \mathbb{Z} is called a *bilateral sequence*. The notations $(x_n)_{n \geq 1}$, $(x_n)_{n \geq 0}$ and $(x_n)_{n \in \mathbb{Z}}$ are used to indicate the index sets \mathbb{P} , \mathbb{N} and \mathbb{Z} , respectively, or simply (x_n) if the intended index set is clear from the context.

1.3.10. *Remark.* Every nonempty set X can be regarded as the range of a family, indexed by $I = X$, with indexing function $x \mapsto x$ (the *identity indexing* of X).

1.3.11. *Remark.* If $f : X \rightarrow Y$ is a function and $A \subset X$, $B \subset Y$ then, by the notations of 1.2.6 and 1.3.1,

$$\begin{aligned} f(A) &= \{y \in Y : xfy \text{ for some } x \in A\} \\ &= \{y \in Y : y = f(x) \text{ for some } x \in A\} = \{f(x) : x \in A\} \end{aligned}$$

$$\begin{aligned} f^{-1}(B) &= \{x \in X : xfy \text{ for some } y \in B\} \\ &= \{x \in X : f(x) = y \text{ for some } y \in B\} = \{x \in X : f(x) \in B\}. \end{aligned}$$

Exercises

1. Discuss $g \circ f$ (including its domain) for arbitrary functions $f : X \rightarrow Y$ and $g : A \rightarrow B$.

{Hint: Consider $f^{-1}(Y \cap A)$.}

2. Consider functions

$$W \xrightarrow{h} X \xrightarrow{f,g} Y \xrightarrow{k} Z.$$

(i) $k \circ (f \circ h) = (k \circ f) \circ h$.

(ii) If $f \circ h = g \circ h$ and h is surjective, then $f = g$.

(iii) If $k \circ f = k \circ g$ and k is injective, then $f = g$.

(iv) If $f \circ h$ is surjective, then so is f .

(v) If $k \circ f$ is injective, then so is f .

(vi) If f and h are injective (surjective, bijective) then so is $h \circ f$.

{Note: If one were to admit an 'empty function' (cf. 1.3.1) then the 'cancellation law' (ii) would fail (contemplate $W = \emptyset$). See also 1.5.5.}

3. If $f : X \rightarrow \mathbb{Z}$ is a function such that $f^2 = f$ (where f^2 is the pointwise product ff), then f is the characteristic function of a suitable subset of X .

4. Let R be relation from X to Y whose domain is X (1.2.10) and let R^{-1} be the reverse of R (1.2.7). Then R is a function from X to Y if and only if $R^{-1}(C \cap D) = R^{-1}(C) \cap R^{-1}(D)$ for every pair of subsets C, D of Y .

1.4. Product Sets, Axiom of Choice

1.4.1. The following two axioms of set theory prepare the way for the next definitions.

Power set axiom: If X is a set, then there exists a set $\mathcal{P}(X)$, called the **power set** of X , whose elements are the subsets of X , that is, $A \in \mathcal{P}(X) \Leftrightarrow A \subset X$.

Union axiom: If \mathcal{A} is a set of sets, then there exists a set X such that $A \subset X$ for all $A \in \mathcal{A}$; the subset U of X defined by

$$U = \{x \in X : x \in A \text{ for some } A \in \mathcal{A}\}$$

then provides a *smallest* set that contains every $A \in \mathcal{A}$, usually written

$$\bigcup \mathcal{A} \text{ or } \bigcup_{A \in \mathcal{A}} A$$

and called the **union** of \mathcal{A} .

1.4.2. *Definition.* If $(A_i)_{i \in I}$ is a family of subsets of a set X , the **union** of the family is the set

$$\bigcup_{i \in I} A_i = \{x \in X : x \in A_i \text{ for some } i \in I\},$$

written briefly $\bigcup A_i$, and the **intersection** of the family is the set

$$\bigcap_{i \in I} A_i = \{x \in X : x \in A_i \text{ for all } i \in I\},$$

written briefly $\bigcap A_i$. Here, the family is a function $I \rightarrow \mathcal{P}(X)$, where $\mathcal{P}(X)$ is the power set of X (1.4.1).

1.4.3. (*De Morgan's formulas*) With the preceding notations,

$$\mathcal{C}(\bigcup A_i) = \bigcap (\mathcal{C}A_i) \quad \text{and} \quad \mathcal{C}(\bigcap A_i) = \bigcup (\mathcal{C}A_i),$$

where \mathcal{C} denotes complement relative to X .

1.4.4. *Definition.* Let \mathcal{A} be a set of sets, $(X_i)_{i \in I}$ a family of sets belonging to \mathcal{A} , that is, a function $I \rightarrow \mathcal{A}$ whose value at $i \in I$ is denoted X_i . Let $U = \bigcup X_i$ (cf. 1.4.1) and contemplate functions $I \rightarrow U$, that is, families $(x_i)_{i \in I}$ of elements of U indexed by I . The **product set** of the family $(X_i)_{i \in I}$ is the set of all families $(x_i)_{i \in I}$ such that $x_i \in X_i$ for all $i \in I$ (in other words, the set of all functions $f : I \rightarrow U$ such that $f(i) \in X_i$ for all $i \in I$); the product set is denoted

$$\prod_{i \in I} X_i,$$

or briefly $\prod X_i$. Two elements $x = (x_i)$ and $y = (y_i)$ of $\prod X_i$ are said to be *equal* if $x_i = y_i$ for all $i \in I$; x_i is called the *i*'th **coordinate** of x , thus two elements x, y of $\prod X_i$ are equal if and only if they are 'coordinatewise equal'.

If $X_j = \emptyset$ for some index j , then no such families (x_i) exist and the product set is empty. That is,

$$X_j = \emptyset \text{ for some } j \Rightarrow \prod_{i \in I} X_i = \emptyset.$$

What if all of the X_i are nonempty? The answer requires a full-fledged axiom of set theory:

1.4.5. **Axiom of choice (AC).** If $(X_i)_{i \in I}$ is a family of nonempty sets then $\prod X_i \neq \emptyset$, that is, there exists a family $(x_i)_{i \in I}$ with $x_i \in X_i$ for all i ; in other words, there exists a function $f : I \rightarrow \bigcup X_i$ such that $f(i) \in X_i$ for all $i \in I$.

1.4.6. The Axiom of Choice is often presented in the following equivalent form:

(AC') If X is a nonempty set, then it is possible to simultaneously select an element from each nonempty subset of X . Stated more formally, there exists a function $c : \mathcal{P}(X) - \{\emptyset\} \rightarrow X$ such that $c(A) \in A$ for every nonempty subset A of X . Such a function c is called a **choice function** for X .

$\{(AC') \Rightarrow (AC)$: With the notations of 1.4.4, let $X = \bigcup X_i$ and let c be a choice function for X ; then $x_i = c(X_i)$ defines an element (x_i) of the product set $\prod X_i$.

$(AC) \Rightarrow (AC')$: If $\mathcal{A} = \mathcal{P}(X) - \{\emptyset\}$, $I = \mathcal{A}$ and \mathcal{A} has the identity indexing $A \mapsto A$ (1.3.10), then any element $c = (c_A)$ of the product set

$$\prod_{A \in \mathcal{A}} A$$

will serve as a choice function for X .)

1.4.7. *Definition.* Let $(X_i)_{i \in I}$ be a family of nonempty sets, $X = \prod X_i$ the product set (nonempty, by the Axiom of Choice). Fix an index $j \in I$. The j 'th **coordinate projection mapping** of X is the mapping $\text{pr}_j : X \rightarrow X_j$ that assigns to each element $x = (x_i)$ of X its j 'th coordinate, that is, $\text{pr}_j(x) = x_j$.

1.4.8. *Theorem.* With notations as in the preceding definition,

(i) for every $j \in I$, the projection mapping $\text{pr}_j : \prod X_i \rightarrow X_j$ is surjective;

(ii) for a pair of mappings $f, g : Y \rightarrow \prod X_i$,

$$f = g \Leftrightarrow \text{pr}_i \circ f = \text{pr}_i \circ g \text{ for all } i \in I.$$

Proof. (i) Let $j \in I$, $a \in X_j$; the problem is to exhibit an element $x = (x_i)$ of $X = \prod X_i$ such that $\text{pr}_j(x) = a$. Write $A_j = \{a\}$ and $A_i = X_i$ for all $i \neq j$; every element of the set $\prod A_i$ (nonempty, by the Axiom of Choice) meets the requirements for x .

(ii) Here $\text{pr}_i \circ f$ denotes the composite mapping $y \mapsto \text{pr}_i(f(y))$ (cf. Figure 6).

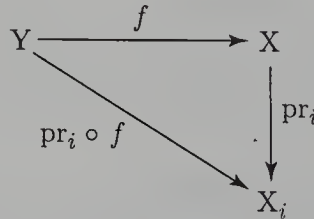


Figure 6

For every $x \in X$, $x = (\text{pr}_i(x))$; thus the condition on the right side of the equivalence of (ii) means that

$$\text{pr}_i(f(y)) = \text{pr}_i(g(y)) \quad \text{for all } y \in Y \text{ and } i \in I,$$

that is,

$$\left(\text{pr}_i(f(y)) \right)_{i \in I} = \left(\text{pr}_i(g(y)) \right)_{i \in I} \quad \text{for all } y \in Y,$$

in other words $f(y) = g(y)$ for all $y \in Y$; that's also what the left side means. \diamond

K. Gödel showed (1939) that the Axiom of Choice is consistent with (i.e., not disprovable from) the 'usual' axioms of set theory, and P. Cohen showed (1963) that it is independent of (i.e., not provable from) the usual axioms. Thus, after a long career as a possible true theorem whose proof had not yet been found—or a possible false theorem for which a counterexample had not yet been constructed—the Axiom of Choice turned out, in fact, to be none of the above; now in semi-retirement as an axiom (lower case) in good standing, its fate reposes on the consistency (freedom from contradiction) of the rest of mathematics.¹ The reader will find an excellent informal essay on these matters in Chapter 3 of I. Kaplansky's *Set theory and metric spaces* [2nd edition, Chelsea, New York, 1977].

In the present text the Axiom of Choice is used, and its use signaled explicitly, whenever (as in 1.4.8) it seems to be the appropriate tool for the discussion at hand.

1.5. Inverse Functions

Recall that if $f : X \rightarrow Y$ is any function and B is a subset of Y , then $f^{-1}(B) = \{x \in X : f(x) \in B\}$ (1.3.11); in this context, f^{-1} is a function that acts on subsets of Y to produce subsets of X , that is, $f^{-1} : \mathcal{P}(Y) \rightarrow \mathcal{P}(X)$. The kind of 'inverse' referred to in the section heading acts instead on points of Y to produce points of X . Whereas the set function $f^{-1} : \mathcal{P}(Y) \rightarrow \mathcal{P}(X)$ exists for every function $f : X \rightarrow Y$, the point function $f^{-1} : Y \rightarrow X$ is defined only when f is bijective (1.5.3 below).

1.5.1. Lemma. *For every function $f : X \rightarrow Y$, there exists a function $g : Y \rightarrow X$ such that $f \circ g \circ f = f$.*

Proof. For each point y of the range $f(X)$ of f , let $A_y = \{x \in X : f(x) = y\} = f^{-1}(\{y\})$; since $y \in f(X)$, A_y is nonempty. We thus have

¹My own view (naive, as cautioned in the chapter introduction) is that the results that have been obtained using the Axiom of Choice are too beautiful to be thrown away without good cause. Innocent until proven guilty; next case . . .

a family $(A_y)_{y \in f(X)}$ of nonempty subsets of X , indexed by the set $f(X)$; by the Axiom of choice, there exists a function $g_0 : f(X) \rightarrow X$ such that $g_0(y) \in A_y$ for all $y \in f(X)$. Choose a point $c \in X$ and let $g : Y \rightarrow X$ be the function defined by the formulas

$$g(y) = \begin{cases} g_0(y) & \text{if } y \in f(X) \\ c & \text{if } y \notin f(X). \end{cases}$$

More precisely, as a set of ordered pairs,

$$g = \{(y, g_0(y)) : y \in f(X)\} \cup \{(y, c) : y \notin f(X)\}.$$

Let $x \in X$; writing $y = f(x) \in f(X)$, we have $g(y) = g_0(y) \in A_y$, therefore

$$f(g(f(x))) = f(g(y)) = f(g_0(y)) = y = f(x),$$

thus g meets the requirements of the lemma. \diamond

1.5.2. Theorem. Let $f : X \rightarrow Y$.

(i) f is surjective \Leftrightarrow there exists a function $g : Y \rightarrow X$ such that $f \circ g = \text{id}_Y$.

(ii) f is injective \Leftrightarrow there exists a function $h : Y \rightarrow X$ such that $h \circ f = \text{id}_X$.

(iii) If f is bijective, necessarily $g = h$.

Proof. (i), \Leftarrow : Assuming g has the indicated properties,

$$(\forall y \in Y) \quad y = \text{id}_Y(y) = (f \circ g)(y) = f(g(y)) \in f(X),$$

whence $f(X) = Y$.

(i), \Rightarrow : By the lemma, there exists a function $g : Y \rightarrow X$ such that $f \circ g \circ f = f$, that is, $(f \circ g) \circ f = \text{id}_Y \circ f$; since f is surjective, it follows that $f \circ g = \text{id}_Y$ (§1.3, Exer. 2).

(ii), \Leftarrow : Suppose that h has the indicated properties. If $f(x) = f(x')$ then $h(f(x)) = h(f(x'))$; since $h \circ f = \text{id}_X$, it follows that $x = x'$.

(ii), \Rightarrow : Let $h : Y \rightarrow X$ be a function such that $f \circ h \circ f = f$, that is, $f \circ (h \circ f) = f \circ \text{id}_X$; since f is injective, it follows that $h \circ f = \text{id}_X$.

(iii) With g and h as in (i) and (ii), $g = \text{id}_X \circ g = (h \circ f) \circ g = h \circ (f \circ g) = h \circ \text{id}_Y = h$. \diamond

1.5.3. Definition. With notations as in 1.5.2, if f is bijective then every g coincides with every h ; the unique function $g : Y \rightarrow X$ such that $g \circ f = \text{id}_X$ and $f \circ g = \text{id}_Y$ is called the **inverse** of f and is denoted f^{-1} .

Since $f^{-1}(f(x)) = x$ ($\forall x \in X$) and $f(f^{-1}(y)) = y$ ($\forall y \in Y$), f^{-1} undoes everything that f does, and vice versa. Some other properties of the inverse are given in the exercises.

1.5.4. Corollary. For sets X and Y ,

$$\exists \text{ injection } X \rightarrow Y \Leftrightarrow \exists \text{ surjection } Y \rightarrow X.$$

Proof. \Rightarrow : If $f : X \rightarrow Y$ is injective, then the mapping $h : Y \rightarrow X$ provided by (ii) of 1.5.2 is surjective (1.3.7).

\Leftarrow : This follows similarly from (i) of 1.5.2 and 1.3.7. \diamond .

1.5.5. *Remark.* Admitting ‘empty functions’ (cf. 1.3.1) would cause trouble in the preceding corollary. If $X = \emptyset$ and $Y \neq \emptyset$ then an injection $\emptyset \rightarrow Y$ makes perfect sense, albeit vacuous (the only subset of $\emptyset \times Y$ available for the graph of a function is \emptyset ; if someone gives us a point in $X = \emptyset$, we stand ready to assign to it a point of Y ; if someone gives us two points of the graph whose first coordinates are distinct, we stand ready to show that the second coordinates are also distinct). However, there exists no function $Y \rightarrow \emptyset$ (if someone gives us a point of Y , there’s nowhere to send it to). To put the matter another way: the empty subset of $\emptyset \times Y$ satisfies conditions (i) and (ii) of Definition 1.3.1, but the empty subset of $Y \times \emptyset$ does not satisfy condition (i). See also §1.3, Exercise 2.

Exercise

1. Let $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ be bijective. Then:

(i) f^{-1} is bijective and $(f^{-1})^{-1} = f$.

(ii) $g \circ f$ is bijective and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

(iii) $f^{-1}(\{y\}) = \{f^{-1}(y)\}$ for all $y \in Y$.

(iv) If B is a subset of Y , then the notation $f^{-1}(B)$ is unambiguous: the inverse image of B under f coincides with the direct image of B under f^{-1} .

1.6. Equivalence Relations, Partitions, Quotient Sets

1.6.1. *Definition.* Let X be a set. A relation R in X (1.2.2) is said to be an **equivalence relation** if it has the following three properties:

(i) xRx for all $x \in X$ (R is *reflexive*);

(ii) if xRy then yRx (R is *symmetric*);

(iii) if xRy and yRz then xRz (R is *transitive*).

Condition (i) says that, as a subset of $X \times X$, R contains the *diagonal* $\Delta = \{(x, x) : x \in X\}$ of $X \times X$. Condition (ii) says that R is stable under reflection in the diagonal: $(x, y) \in R \Rightarrow (y, x) \in R$. The graphical interpretation of (iii) is less edifying.

1.6.2. *Examples.* (1) For $x, y \in \mathbb{Z}$, write xRy if $x - y$ is divisible by 6; this equivalence relation is usually written $x \equiv y \pmod{6}$, read “ x is congruent to y modulo 6”.

(2) Let $f : X \rightarrow Y$ be any function. For $x, x' \in X$, write xRx' if $f(x) = f(x')$; R is called the equivalence relation *deduced* from f .

(3) The diagonal $\Delta = \{(x, x) : x \in X\}$ of $X \times X$ is the equivalence relation of equality in X .

(4) The set $X \times X$ is an equivalence relation in X , called the *trivial relation* (because xRy for all x, y in X).

(5) Let X be the set of all lines in 3-space. For $l, m \in X$ write lRm if l is either equal to or parallel to m .

Some popular notations for equivalence relations: $=, \equiv, \sim, \cong, \approx$.

1.6.3. *Definition.* Let X be a set equipped with an equivalence relation \sim and let $x \in X$. The **equivalence class** of x for \sim is the set $[x]$ of all elements of X that are equivalent to x , that is,

$$[x] = \{y \in X : y \sim x\}.$$

{Other convenient notations for the equivalence class of x are \dot{x} , \bar{x} , x' .}

1.6.4. *Examples.* For the examples of 1.6.2, (1) $[x] = \{x + 6k : k \in \mathbb{Z}\}$, (2) $[x] = f^{-1}(\{f(x)\}) = \{x' \in X : f(x') = f(x)\}$, (3) $[x] = \{x\}$, (4) $[x] = X$, and (5) $[l]$ is the set of all lines having a common direction.

1.6.5. *Theorem.* Let \sim be an equivalence relation in X . With notations as in Definition 1.6.3,

(i) $x \in [x]$ for all $x \in X$; thus every equivalence class $[x]$ is nonempty and X is the union of all the equivalence classes:

$$X = \bigcup_{x \in X} [x].$$

(ii) The equivalence classes are pairwise disjoint; that is, if $x, y \in X$ then either $[x] = [y]$ or $[x] \cap [y] = \emptyset$ (but not both).

(iii) $x \sim y \Leftrightarrow [x] = [y]$.

(iv) $x \not\sim y \Leftrightarrow [x] \cap [y] = \emptyset$.

Proof. (i) $x \in [x]$ because $x \sim x$ (reflexivity).

(ii) It suffices to show that if the sets $[x]$ and $[y]$ have a point in common then they are identical. Say $z \in [x] \cap [y]$; thus $z \sim x$ and $z \sim y$, therefore $x \sim z$ (symmetry) and, since $x \sim z$ and $z \sim y$, also $x \sim y$ (transitivity). If $t \in [x]$ then $t \sim x$; also $x \sim y$, therefore $t \sim y$, thus $t \in [y]$. This shows that $[x] \subset [y]$, and similarly $[y] \subset [x]$.

(iii) If $x \sim y$ then $x \in [x] \cap [y]$, therefore $[x] = [y]$ by (ii). Conversely, if $[x] = [y]$ then $x \in [x] = [y]$, therefore $x \sim y$.

(iv) The condition on the left is the negation of $x \sim y$, whereas that on the right is (in view of (ii)) the negation of $[x] = [y]$, thus (iv) is just the contrapositive form of (iii). \diamond

1.6.6. *Definition.* A **partition** of a set X is a set \mathcal{A} of pairwise disjoint, nonempty subsets of X whose union is X , that is, (i) $A \neq \emptyset$ ($\forall A \in \mathcal{A}$), (ii) $A, B \in \mathcal{A}$, $A \neq B \Rightarrow A \cap B = \emptyset$, and (iii) $\bigcup \mathcal{A} = X$.

According to 1.6.5, the equivalence classes for an equivalence relation on X constitute a partitioning of X . Conversely,

1.6.7. Theorem. Let \mathcal{A} be a partition of a set X . For $x, y \in X$, write $x \sim y$ if there exists a set $A \in \mathcal{A}$ containing both x and y . Then \sim is an equivalence relation in X for which the set of equivalence classes is \mathcal{A} .

The proof of 1.6.7 is straightforward and elementary.

1.6.8. Definition. Let \sim be an equivalence relation in X . The set of equivalence classes for \sim is called the **quotient set** of X by the relation \sim and is denoted X/\sim ; thus,

$$X/\sim = \{[x] : x \in X\} \subset \mathcal{P}(X).$$

Assuming X is nonempty, the mapping $q : X \rightarrow X/\sim$ defined by $q(x) = [x]$ is called the **quotient mapping** for the relation.

1.6.9. Examples. For the examples of 1.6.2 and 1.6.4, (1) $\mathbb{Z}/R = \mathbb{Z}_6$ (the set of ‘integers modulo 6’), (2) the elements of X/R are the sets of constancy of f , (3) the quotient mapping $X \rightarrow X/R$ is the bijection $x \mapsto \{x\}$, (4) X/R consists of a single element (and the quotient mapping is constant), and (5) an element $[l]$ of the quotient set can be identified with a (two-way) ‘direction’.

1.6.10. Theorem. If X is a nonempty set and \sim is an equivalence relation in X , then the quotient mapping $q : X \rightarrow X/\sim$ is surjective and \sim coincides with the equivalence relation deduced from q (1.6.2).

Proof. Straightforward. \diamond

The essential message of the foregoing theorems is that equivalence relations, partitions, and sets of constancy of a function are three presentations of the same basic concept.

1.6.11. Let $f : X \rightarrow Y$ be any function, \sim the equivalence relation deduced from f (1.6.2). Write \dot{x} for the equivalence class of $x \in X$, $\dot{X} = X/\sim$ for the quotient set, $q : X \rightarrow \dot{X}$ for the quotient mapping, and $i : f(X) \rightarrow Y$ for the insertion mapping $i = i_{f(X)}$ (1.3.4); in particular, q is surjective and i is injective (cf. Figure 7; the lower arrow anticipates the next theorem).

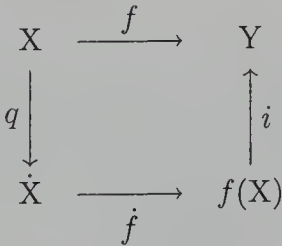


Figure 7

1.6.12. Theorem. *With the preceding notations, there exists a unique bijection $\dot{f} : \dot{X} \rightarrow f(X)$ such that $f = i \circ \dot{f} \circ q$.*

Proof. Uniqueness: If $g, h : \dot{X} \rightarrow f(X)$ are functions such that $i \circ g \circ q = i \circ h \circ q$ then, citing §1.3, Exer. 2, $g \circ q = h \circ q$ (because i is injective) and $g = h$ (because q is surjective).

Existence: Let $u \in X$; the problem is to define $\dot{f}(u)$. If $u = q(x) = q(y)$ then $x \sim y$, that is, $f(x) = f(y)$, therefore the formula $\dot{f}(u) = f(x)$ defines a function $\dot{f} : X \rightarrow f(X)$ unambiguously. It is clear from the definitions that $i(\dot{f}(q(x))) = i(f(x)) = f(x)$, and easy to check that \dot{f} is bijective. $\{\dot{f}$ is the inverse of the mapping $z \mapsto f^{-1}(\{z\})$ ($z \in f(X)$). $\}$ \diamond

The formula $f = i \circ \dot{f} \circ q$ is called the *canonical factorization* of f (as the composite of a surjection, followed by a bijection, followed by an injection).

Exercise

1. Let X and Y be sets, R a relation in X and S a relation in Y . A mapping $f : X \rightarrow Y$ is said to be *compatible* with R and S in case (for x, x' in X) $xRx' \Rightarrow f(x)Sf(x')$.

If R and S are equivalence relations (1.6.1) in X and Y and if $f : X \rightarrow Y$ is compatible with R and S , then there exists a unique mapping $g : X/R \rightarrow Y/S$ such that

$$(*) \quad g \circ Q_X = Q_Y \circ f,$$

where $Q_X : X \rightarrow X/R$ and $Q_Y : Y \rightarrow Y/S$ are the quotient mappings (1.6.8). The relation (*) is expressed by saying that the following diagram is *commutative* (the two ways of getting from X to Y/S are equal):

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ Q_X \downarrow & & \downarrow Q_Y \\ X/R & \xrightarrow{g} & Y/S \end{array}$$

1.7. Order Relations

1.7.1. *Definition.* A relation \leq in a set X is said to be a **partial ordering** of X (or to be an *order relation* in X) if it is reflexive, transitive and 'antisymmetric':

- (i) $x \leq x$ for all $x \in X$ (*reflexivity*),
- (ii) $x \leq y$ & $y \leq z \Rightarrow x \leq z$ (*transitivity*),

(iii) $x \leq y \ \& \ y \leq x \Rightarrow x = y$ (*antisymmetry*).

A relation satisfying only (i) and (ii) is called a **pre-ordering** of X . Other notations in use for such relations are \geq , \subset , \supset , \prec , \succ , etc.

A **simple ordering** (or *total ordering*, *linear ordering*) is a partial ordering \leq for which any two elements are comparable, that is,

(iv) $(\forall x, y) \ x \leq y \text{ or } y \leq x$.

A **pre-ordered set** is a pair (X, \leq) , where \leq is a pre-ordering of X ; the terms **partially ordered set** and **simply ordered set** are defined similarly.

1.7.2. *Examples.* (1) In the set of all positive integers, the relation $x|y$ (x divides y) is a partial ordering (but not a simple ordering).

(2) The relation $A \subset B$ is a partial ordering of the set $\mathcal{P}(X)$ of all subsets of a set X (not a simple ordering when X has more than one element).

(3) The usual relation \leq in the set of all positive integers is a simple ordering.

(4) For every set X , the relation in X defined by $X \times X$ is a pre-ordering of X ($x \leq x'$ for all $x, x' \in X$), called the *trivial pre-ordering* of X . When X has more than one element, the trivial pre-ordering is not a partial ordering.

(5) Every equivalence relation (1.6.1) in a set X is a pre-ordering; the only equivalence relation in X that is a partial ordering is the identity relation (the relation \leq satisfying $x \leq x$ for all $x \in X$, and nothing else), defined by the diagonal Δ of $X \times X$.

1.7.3. *Example.* If (X, \leq) is a pre-ordered set and A is a subset of X , the relation on A induced by \leq (cf. 1.2.5) is a pre-ordering of A . If the relation on X is a partial ordering (simple ordering), then so is the relation induced on A . A simply ordered subset of X is called a **chain** in X . For example, if X is partially ordered, $A = \{x_1, \dots, x_n\} \subset X$ and $x_1 \leq x_2 \leq \dots \leq x_n$, then A is a chain.

1.7.4. *Example.* In a pre-ordered set (X, \leq) , write $x \leq' y$ in case $y \leq x$ (\leq' is the 'reverse' of \leq in the sense of 1.2.7); the relation \leq' is also a pre-ordering (called the *dual* of \leq). When \leq is a partial order (simple order), then so is \leq' .

1.7.5. *Notations.* In a pre-ordered set (X, \leq) , $y \geq x$ is an alternative notation for $x \leq y$. If $x \leq y$ and $x \neq y$, we write $x < y$ (or $y > x$).

{CAUTION: If X is a set equipped with the trivial pre-ordering (1.7.2), then $a < b$ and $b < a$ for every pair a, b of distinct elements of X .}

When \leq is a simple ordering, exactly one of the statements

$$x < y, \ x = y, \ x > y$$

is true for any given pair of elements x, y (*Law of trichotomy*).

1.7.6. *Example.* If (X, \leq) is a partially ordered set and (Y, \leq) is any pre-ordered set, then the relation $(x, y) \leq (x', y')$ in $X \times Y$ defined by the condition

$$\ll \text{either } x < x', \text{ or } x = x' \ \& \ y \leq y' \gg$$

is a pre-ordering of $X \times Y$, called the *lexicographic* pre-ordering (as in a dictionary of two-letter words). If both X and Y are partially ordered (simply ordered) then so is $X \times Y$ for the lexicographic pre-ordering.

Another pre-ordering of $X \times Y$ is defined by the condition

$$\ll x \leq x' \ \& \ y \leq y' \gg;$$

it is called the *product* of the given pre-orderings. If X and Y are partially ordered then so is $X \times Y$, but the analogous statement for simple ordering is false.

1.7.7. The study of pre-orderings is effectively reduced to the study of partial orderings by 'passing to quotients' in an appropriate way. Suppose (X, \leq) is a pre-ordered set. For $x, y \in X$, write $x \sim y$ in case both $x \leq y$ and $y \leq x$. It is easy to see that \sim is an equivalence relation on X . If u, v are elements of the quotient set X/\sim , say $u = [x]$ and $v = [y]$, we propose to define a relation $u \leq v$ in case $x \leq y$; if also $u = [x']$ and $v = [y']$, then $x \leq y \Leftrightarrow x' \leq y'$, so the definition is unambiguous. The relation so defined in X/\sim is easily seen to be a partial ordering (called the *quotient order relation*); for example, if $u \leq v$ and $v \leq u$ then, with the preceding notations, $x \leq y$ and $y \leq x$, therefore $x \sim y$ and so $u = v$.

The next concepts pertain to relations between elements and subsets of an ordered set.

1.7.8. *Definition.* Let (X, \leq) be a pre-ordered set, $A \subset X$ and $c \in X$.

(i) If $x \leq c$ ($\forall x \in A$), we say that c is an **upper bound** for A (or that c *majorizes* A , or that c is a *majorant* of A). Dually,

(ii) If $c \leq x$ ($\forall x \in A$), we say that c is a **lower bound** for A (or that c *minorizes* A , or that c is a *minorant* of A).

If A has a majorant (minorant) in X , it is said to be *bounded above* (*bounded below*); A is said to be *bounded* if it is both bounded above and bounded below.

1.7.9. *Remark.* Suppose (X, \leq) is a partially ordered set and $a \in A \subset X$. If a majorizes (minorizes) A then it is uniquely determined by this property; it is called the *largest* (*smallest*) element of A . The following concept is more subtle:

1.7.10. *Definition.* Let (X, \leq) be a pre-ordered set, $A \subset X$. An element $a \in A$ is said to be a *maximal* element of A if

$$(x \in A \ \& \ x \geq a) \Rightarrow x = a$$

(in other words, A contains no element $> a$); dually, if $a \in A$ satisfies the condition

$$(x \in A \ \& \ x \leq a) \Rightarrow x = a$$

then a is said to be a *minimal* element of A .

1.7.11. *Remark.* In a partially ordered set (X, \leq) , if $a \in A$ is the largest element of A then a is maximal in A . If (X, \leq) is simply ordered and if $a \in A$ is maximal in A , then a is the largest element of A . Similarly for “smallest” and “minimal”. Thus, when X is simply ordered, the concepts of maximal element and largest element coincide (as do the concepts of minimal element and smallest element).

1.7.12. *Examples.* (1) In the field \mathbb{Q} of rational numbers, with the usual ordering, the set $A = \{r \in \mathbb{Q} : 0 < r < 1\}$ is bounded, but has neither a largest nor a smallest element.

(2) For the usual ordering of \mathbb{P} , every nonempty subset has a smallest element (*Principle of mathematical induction*).

(3) Let \mathcal{S} be the set of all nonempty subsets A of \mathbb{P} such that A has at most 5 elements. Order \mathcal{S} by the inclusion relation \subset . Every 5-element subset of \mathbb{P} is a maximal element of \mathcal{S} , and every singleton in \mathbb{P} is a minimal element of \mathcal{S} .

The rest of the section prepares the way for the discussion of well-ordered sets in §1.14; it can be deferred until then.

1.7.13. *Definition.* Let (X, \leq) and (Y, \leq) be pre-ordered sets, $f : X \rightarrow Y$ a function. We say that f is (1) an **order morphism** if $x \leq x' \Rightarrow f(x) \leq f(x')$, (2) an **order isomorphism** if f is bijective and both f and f^{-1} are order morphisms, and (3) an **order monomorphism** if f is injective and $x \leq x' \Leftrightarrow f(x) \leq f(x')$. {For a possible definition of ‘order epimorphism’, see (iii) of Exercise 6.}

Condition (1) says that f is compatible with the order relations in the sense of §1.6, Exercise 1. Condition (2) says that f is bijective and $x \leq x' \Leftrightarrow f(x) \leq f(x')$. Condition (3) says that f is injective and the bijection $X \rightarrow f(X)$ having the same graph as f is an order isomorphism of X onto the set $f(X)$ equipped with the pre-ordering it inherits from Y .

1.7.14. *Definition.* Pre-ordered sets X and Y are said to be **similar**, written $X \approx Y$, if there exists an order isomorphism $X \rightarrow Y$; if X and Y are *not* similar, we write $X \not\approx Y$. Convention: $\emptyset \approx \emptyset$.

1.7.15. *Remarks.* (i) In every set of pre-ordered sets, similarity is an equivalence relation: $X \approx X$, $X \approx Y \Rightarrow Y \approx X$, and $(X \approx Y \ \& \ Y \approx Z) \Rightarrow X \approx Z$.

(ii) In the set \mathbb{P} of positive integers, the relations $m|n$ (m divides n) and $m \leq n$ (the usual relation) are partial orderings. The identity mapping $f : (\mathbb{P}, |) \rightarrow (\mathbb{P}, \leq)$ is an injective order morphism, but it is not an order monomorphism.

(iii) Consider the subsets $X = (0, 2)$ and $Y = (0, 1] \cup [3/2, 2)$ of the real line, each equipped with the usual (simple) ordering. There exist order monomorphisms $X \rightarrow Y$ and $Y \rightarrow X$ (for example, $x \mapsto \frac{1}{2}x$ and $x \mapsto x$, respectively), but $X \not\approx Y$ (Y has a pair of points with nothing in between, whereas X does not).

1.7.16. Proposition. *If X is a partially ordered set, Y is a pre-ordered set, and $f : X \rightarrow Y$ satisfies $x \leq x' \Leftrightarrow f(x) \leq f(x')$, then f is an order monomorphism.*

Proof. We need only show that f is injective. If $f(x) = f(x')$ then $f(x) \leq f(x')$ and $f(x') \leq f(x)$, therefore $x \leq x'$ and $x' \leq x$; since X is partially ordered, $x = x'$. \diamond

1.7.17. Proposition. *If X is a simply ordered set, Y is a partially ordered set, and $f : X \rightarrow Y$ is an injective order morphism, then f is an order monomorphism.*

Proof. Assuming $f(x) \leq f(x')$, we must show that $x \leq x'$. The alternative is that $x' \leq x$ and $x \neq x'$; then $f(x') \leq f(x)$, whence $f(x) = f(x')$ (Y is partially ordered), contrary to the injectivity of f . \diamond

Exercises

1. In the set of all sequences $s = (a_n)$, $t = (b_n)$, ... of positive integers, the relation $s \leq t$ defined by $\ll a_n \leq b_n \ (\forall n) \gg$ is a partial ordering (but not a simple ordering).

2. If X is a set, (Y, \leq) is a partially ordered set, and $\mathcal{F} = \mathcal{F}(X, Y)$ is the set of all functions $f : X \rightarrow Y$, $g : X \rightarrow Y$, ..., then the relation $f \leq g$ defined by $\ll f(x) \leq g(x) \ (\forall x \in X) \gg$ is a partial ordering of \mathcal{F} .

3. Let $(X_i, \leq_i)_{i \in I}$ be a family of partially ordered sets, $X = \prod X_i$ the product set. The relation $(x_i) \leq (y_i)$ in X defined by $\ll x_i \leq_i y_i \ (\forall i \in I) \gg$ is a partial ordering of X (called the *product ordering*).

4. (i) Let X be a set and let \leq be a relation in X , that is, a subset of $X \times X$ (1.2.2). {For the moment, no properties of \leq are assumed; in particular, it need not be a pre-order relation.} As in 1.7.5, define $x < y$ to mean that $x \leq y$ and $x \neq y$. (If \leq is the empty relation or if X has only one element, we have not defined anything.) Then $x \leq y \Rightarrow x = y$ or $x < y$, and the reverse implication holds when the relation is reflexive.

(ii) If X and Y are sets, each with a relation \leq as in (i), we may define the 'morphism' concepts exactly as in 1.7.13. In particular, a bijection

$f : X \rightarrow Y$ is called an *isomorphism* for the relations if $x \leq y \Leftrightarrow f(x) \leq f(y)$. Remark (i) of 1.7.15 remains valid in the present context.

(iii) With X and Y as in (ii), let $Z = X \times Y$ be the product set. If $z = (x, y)$ and $z' = (x', y')$ are points of Z , define $z \leq z'$ to mean that either (1) $z = z'$, or (2) $x < x'$, or (3) $x = x'$ and $y < y'$. (This relation in Z is by definition reflexive.) Conditions (1)–(3) say that either $z = z'$ or, if $z \neq z'$ then in the first coordinate in which z and z' differ, the coordinate of z is \leq the coordinate of z' . If the given relation on Y is reflexive, then the relation so defined on Z coincides with the relation defined by the condition in 1.7.6.

(iv) Let X_1, \dots, X_n be sets, let $X = X_1 \times \dots \times X_n$ be the product set, and suppose that for each index i we have a relation in X_i , denoted \leq for simplicity. Define a relation \leq in X as follows: given $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, define $x \leq y$ to mean that either (1) $x = y$, or (2) $x \neq y$ and $x_j < y_j$ for the first index j such that $x_j \neq y_j$. We call this the *lexicographic* relation in X derived from the relations in the X_i . When $n = 2$ and the relation in X_2 is reflexive, this relation on X coincides with the relation defined by the condition in 1.7.6.

(v) With notations as in (iv), the natural bijection $(X_1 \times \dots \times X_{n-1}) \times X_n \rightarrow X$ is an isomorphism (in the sense of (ii)) for the lexicographic relations, therefore

$$(X_1 \times \dots \times X_{n-1}) \times X_n \approx X_1 \times \dots \times X_n.$$

More generally, $(X_1 \times \dots \times X_{k-1}) \times (X_k \times \dots \times X_n) \approx X_1 \times \dots \times X_n$ for every index k with $1 < k \leq n$.

{Hint: To make it interesting, assume $n \geq 3$. Let $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ be points of the right side, $x' = ((x_1, \dots, x_{n-1}), x_n)$, $y' = ((y_1, \dots, y_{n-1}), y_n)$ the corresponding points of the left side. If $x' \leq y'$ then either $x' = y'$, or $(x_1, \dots, x_{n-1}) = (y_1, \dots, y_{n-1})$ and $x_n < y_n$, or $(x_1, \dots, x_{n-1}) < (y_1, \dots, y_{n-1})$; in the last case, $x_j < y_j$ for the first index j such that $x_j \neq y_j$. In all three cases, $x \leq y$. That's half the battle.}

(vi) With notations as in (iv), $(X_1 \times X_2) \times X_3 \approx X_1 \times (X_2 \times X_3)$ via the natural bijection.

(vii) With notations as in (iv), if all of the X_i are partially ordered (simply ordered) then so is X . {Hint: $n = 2$ and induction.}

5. If $[a, b]$ is a closed interval in \mathbb{R} , and \mathcal{S} is the set of all subdivisions $\sigma = \{a = x_0 < x_1 < \dots < x_n = b\}$, $\tau = \{a = y_0 < y_1 < \dots < y_m = b\}$, ... of $[a, b]$, then the relation $\sigma \succ \tau$ defined by $\ll \sigma$ is a refinement of $\tau \gg$ (every y_i is some x_j) is a partial ordering of \mathcal{S} . (Two subdivisions are regarded as being 'equal' if they are specified by the same points intermediate to a and b , in other words, both $\sigma \succ \tau$ and $\tau \succ \sigma$.)

6. Let (X, \leq) and (Y, \leq) be pre-ordered sets and let $(X/\sim, \leq)$ and $(Y/\sim, \leq)$ be the partially ordered sets derived from them by the technique of 1.7.7; in particular, for $x, x' \in X$, $x \sim x'$ means that $x \leq x'$ and $x' \leq x$, and $[x]$ denotes the equivalence class of $x \in X$ for the relation \sim . Let $f: X \rightarrow Y$.

(i) If f is an order morphism and $g: X/\sim \rightarrow Y/\sim$ is defined by $g([x]) = [f(x)]$ for all $x \in X$ (cf. §1.6, Exercise 1), then g is an order morphism.

(ii) If $x \leq x' \Leftrightarrow f(x) \leq f(x')$, then g is an order monomorphism.

(iii) If f is surjective and $x \leq x' \Leftrightarrow f(x) \leq f(x')$, then g is an order isomorphism.

7. If X is not partially ordered, the assertion of 1.7.16 may be false. {Hint: Let $X = \{x_1, x_2\}$, $x_1 \neq x_2$, equipped with the trivial pre-ordering $x_i \leq x_j$ for all i and j , and let $Y = \{y\}$, equipped with the only available pre-ordering.}

8. If X is a set containing more than one element and equipped with the trivial pre-ordering (1.7.2, (4)) and Y is a set equipped with a pre-ordering that is not the trivial pre-ordering, then the relation in $X \times Y$ defined by the condition in 1.7.6 is not a pre-ordering (it is not transitive).

{Hint: X contains elements x, x' with $x < x'$ and $x' < x$, whereas Y contains elements y, y' for which $y \leq y'$ does not hold. Contemplate the points (x, y) , (x', y) , (x, y') of $X \times Y$.}

1.8. Real Numbers

In a first course in real analysis, the starting point is often a set of axioms for the field \mathbb{R} of real numbers, as a **complete ordered field**.¹ These consist in the purely algebraic ‘field axioms’ (properties of addition and multiplication), axioms for the set of ‘positive’ elements (the basis for a simple ordering of \mathbb{R}), and the decisive axiom that distinguishes \mathbb{R} from all other ‘ordered fields’:

Completeness axiom: *Every nonempty subset of \mathbb{R} that is bounded above has a smallest majorant (that is, a ‘least upper bound’).*

It is relatively easy to show that such a field is *unique*, in the sense that any two such fields are isomorphic. (First establish an isomorphism between their ‘rational subfields’—the subfields generated by their respective unity elements—then use the order-density of the rationals² to extend the isomorphism.) In the present section, we sketch a proof of the *existence* of a complete ordered field, starting from the field \mathbb{Q} of rational numbers (on the grounds that, time permitting, the less one takes on faith, the better).

¹ Cf. the author, Chapter 1 of *A first course in real analysis* [Springer-Verlag, New York, 1994], henceforth cited briefly as *First course*.

² *First course*, Theorem 2.4.1.

The existence of a complete ordered field was first demonstrated by R. Dedekind (ca. 1858), by a method now known as ‘Dedekind cuts’.³ G. Cantor subsequently gave a construction based on Cauchy sequences.⁴ Roughly speaking, the advantage of Dedekind’s method is that it is applicable to more general ordered structures; the virtues of Cantor’s method are greater ease in extending the algebraic operations and its applicability to more general ‘uniform structures’.⁵ The method to be sketched here is that of Cantor.

1.8.1. *Definition.* With the usual ordering of the field \mathbb{Q} of rational numbers, the **absolute value** function on \mathbb{Q} is defined by the formulas

$$|r| = \begin{cases} r & \text{if } r \geq 0 \\ -r & \text{if } r < 0. \end{cases}$$

1.8.2. *Lemma.* Let $r, s \in \mathbb{Q}$.

- (1) $|r| = 0 \Leftrightarrow r = 0$; $|r| > 0 \Leftrightarrow r \neq 0$.
- (2) $|rs| = |r||s|$.
- (3) If $s \geq 0$, then $|r| \leq s \Leftrightarrow -s \leq r \leq s$.
- (4) $|r + s| \leq |r| + |s|$ (*Triangle inequality*).

Proof. (3) From $-s \leq r \leq s$ one infers both $-r \leq s$ and $r \leq s$; one of $r, -r$ is $|r|$, thus $|r| \leq s$. The reverse implication follows from the fact that both $r \leq |r|$ and $-r \leq |r|$.

(4) Note that $-(|r| + |s|) \leq r + s \leq |r| + |s|$ and cite (3). \diamond

1.8.3. *Remark.* If $|r| \leq s$ for all $s > 0$ (here $r, s \in \mathbb{Q}$), then $r = 0$. {If $r \neq 0$ consider $s = \frac{1}{2}|r|$.}

1.8.4. *Definition.* A sequence (r_n) in \mathbb{Q} is said to be **bounded** if the set $\{r_n : n \in \mathbb{P}\}$ is bounded in the sense of 1.7.8. The set of all bounded sequences in \mathbb{Q} will be denoted \mathcal{B} .

1.8.5. *Remark.* A subset S of \mathbb{Q} is bounded if and only if there exists a positive $t \in \mathbb{Q}$ such that $|s| \leq t$ for all $s \in S$.

{Proof: Suppose S is bounded, that is, $a \leq s \leq b$ ($\forall s \in S$) for suitable $a, b \in \mathbb{Q}$; since $-|a| \leq a$ and $b \leq |b|$, the larger of $|a|$ and $|b|$ meets the requirements for t (as does $|a| + |b|$). If, conversely, $|s| \leq t$ for all $s \in S$, then $-t \leq s \leq t$ ($\forall s \in S$) shows that S is bounded.}

³ Cf. R. Dedekind, *Essays on the theory of numbers* [Translated from the German original, Open Court Publ. Co., LaSalle, 1901; reprinted by Dover, New York], E. Landau, *Foundations of analysis* [Chelsea, New York, 1951].

⁴ Cf. E. W. Hobson, *The theory of functions of a real variable and the theory of Fourier series*, vol. 1, p. 28 [Dover, New York, 1957], E. Hewitt and K. Stromberg, *Real and abstract analysis*, §5 [Springer-Verlag, New York, 1965].

⁵ N. Bourbaki, *General topology* [Addison-Wesley, Reading, Mass., 1966], Chapter II, §3.

In particular, a sequence (r_n) in \mathbb{Q} is bounded if and only if $|r_n| \leq t$ ($\forall n$) for some positive $t \in \mathbb{Q}$.

1.8.6. Lemma. *The set \mathcal{B} of all bounded sequences in \mathbb{Q} is a commutative ring with unity for the term-by-term operations*

$$(r_n) + (s_n) = (r_n + s_n), \quad (r_n)(s_n) = (r_n s_n),$$

with unity element the constant sequence (1) all of whose terms are equal to 1.

Proof. \mathcal{B} is closed under these operations by the relations $|r_n + s_n| \leq |r_n| + |s_n|$ and $|r_n s_n| = |r_n| |s_n|$. \diamond

1.8.7. Definition. A sequence (r_n) in \mathbb{Q} is said to be **Cauchy** if, for every $t > 0$ in \mathbb{Q} , there exists an index N such that

$$m, n \geq N \Rightarrow |r_m - r_n| \leq t.$$

This is also expressed by saying that

$$(\forall t \in \mathbb{Q}, t > 0) \quad |r_m - r_n| \leq t \text{ ultimately.}$$

We write \mathcal{C} for the set of all Cauchy sequences in \mathbb{Q} .

1.8.8. Lemma. \mathcal{C} is a subring of \mathcal{B} containing the constant sequences.

Proof. The crux of the matter is to show that every Cauchy sequence (r_n) is bounded; this follows from the fact that $|r_m - r_n| \leq 1$ from some index onward—say for $m, n \geq N$ —and the inequality $|r_n| \leq |r_n - r_N| + |r_N|$. If (r_n) and (s_n) are both Cauchy, then the identity

$$r_m s_m - r_n s_n = r_m (s_m - s_n) + (r_m - r_n) s_n$$

shows that their product $(r_n s_n)$ is also Cauchy. The closure of \mathcal{C} under addition follows at once from the triangle inequality (1.8.2). \diamond

1.8.9. Definition. A sequence (z_n) in \mathbb{Q} is said to be **null** if

$$(\forall t \in \mathbb{Q}, t > 0) \quad |z_n| \leq t \text{ ultimately;}$$

that is, for every rational $t > 0$, there exists an index N such that $n \geq N \Rightarrow |z_n| \leq t$. We write \mathcal{N} for the set of all null sequences in \mathbb{Q} .

1.8.10. Example. The sequence $(1/n)$ is null. {Proof: If $t = M/N$ ($M, N \in \mathbb{P}$) then the relation $1/n \leq t$ (that is, $N \leq Mn$) holds, for example, for all $n \geq N$.}

1.8.11. Lemma. \mathcal{N} is an ideal of \mathcal{C} (and of \mathcal{B}).

Proof. It is obvious from the triangle inequality that \mathcal{N} is an additive subgroup of \mathcal{C} . If (b_n) is bounded and (z_n) is null, it follows from $|b_n z_n| = |b_n| |z_n|$ that $(b_n z_n)$ is null. \diamond

1.8.12. *Definition.* We write $\mathbb{R} = \mathcal{C}/\mathcal{N}$ for the quotient ring of \mathcal{C} modulo \mathcal{N} , and $(r_n)' = (r_n) + \mathcal{N}$ for the coset of $(r_n) \in \mathcal{C}$; thus $(r_n) \mapsto (r_n)'$ is the quotient mapping (a homomorphism of \mathcal{C} onto \mathbb{R}).

Our task is to show that \mathbb{R} is a complete ordered field. At any rate, it is elementary that \mathbb{R} is a commutative ring with unity element $(1) + \mathcal{N}$.

If $r \in \mathbb{Q}$ and (r) is the constant sequence with all terms equal to r , we abbreviate $(r)' = (r) + \mathcal{N}$ to r' ; thus, $r \mapsto r'$ is a mapping $\mathbb{Q} \rightarrow \mathbb{R}$.

1.8.13. *Lemma.* The mapping $r \mapsto r'$ ($r \in \mathbb{Q}$) is a ring monomorphism $\mathbb{Q} \rightarrow \mathbb{R}$.

Proof. The mapping $r \mapsto r'$ is the composite of the homomorphism $r \mapsto (r)$ of \mathbb{Q} into \mathcal{C} with the quotient homomorphism $\mathcal{C} \rightarrow \mathbb{R}$. If $r' = 0'$ then (r) is null, therefore $r = 0$ (1.8.3), whence injectivity. \diamond

1.8.14. *Remark.* If (r_n) is a Cauchy sequence in \mathbb{Q} , N is a positive integer, and (s_n) is a sequence in \mathbb{Q} such that $s_n = r_n$ for all $n > N$, then (s_n) is also Cauchy and $(r_n)' = (s_n)'$; in other words, one can modify r_k for $k = 1, \dots, N$ without changing the element $(r_n)'$ of \mathbb{R} .

1.8.15. *Lemma.* \mathbb{R} is a field.

Proof. Let $x \in \mathbb{R}$, $x \neq 0$; we seek an element $y \in \mathbb{R}$ such that $xy = 1$ (more precisely, $1'$). Say $x = (r_n)'$. Since $x \neq 0$, (r_n) is not a null sequence, thus there exists a rational $t > 0$ such that $|r_n|$ fails to be ultimately $< t$. This means that $|r_n| \geq t$ 'frequently', that is, $|r_{n_k}| \geq t$ for a sequence of indices $n_1 < n_2 < n_3 < \dots$.

We assert that $|r_n| \geq t/2$ ultimately. For, since (r_n) is Cauchy, there exists an index N such that $|r_m - r_n| \leq t/2$ for all $m, n \geq N$; if k is an index such that $n_k \geq N$ then, for all $n \geq N$,

$$t \leq |r_{n_k}| \leq |r_{n_k} - r_n| + |r_n| \leq t/2 + |r_n|,$$

whence $|r_n| \geq t/2$.

We are ready to define the required element y . Let (s_n) be the sequence in \mathbb{Q} defined by

$$s_n = \begin{cases} 0 & \text{for } n < N \\ 1/r_n & \text{for } n \geq N. \end{cases}$$

From the preceding remark, we see that $|s_n| \leq 2/t$ for all n , so (s_n) is bounded. In fact, (s_n) is Cauchy. For, if $m, n \geq N$ then

$$|s_m - s_n| = \frac{1}{|r_m| |r_n|} \cdot |r_n - r_m| \leq \frac{4}{t^2} \cdot |r_n - r_m|;$$

for every rational $r > 0$, $|r_n - r_m| \leq (t^2/4)r$ (and therefore $|s_m - s_n| \leq r$) for m and n sufficiently large. Let $y = (s_n)'$. Since $r_n s_n = 1$ for all $n \geq N$, it follows that $xy = 1$. \diamond

So far, everything is relatively straightforward; the order relation in \mathbb{R} is more delicate.

1.8.16. *Definition.* For $x \in \mathbb{R}$, we write $x \geq 0$ in case $x = (r_n)'$ with $r_n \geq 0$ for all n . {It is the same to require that $x = (r_n)'$ with $r_n \geq 0$ ultimately.}

1.8.17. *Remarks.* (1) If $x, y \in \mathbb{R}$ and $x \geq 0$, $y \geq 0$, then also $x + y \geq 0$ and $xy \geq 0$.

(2) If $x \geq 0$ and $-x \geq 0$ then $x = 0$. {Proof: If $x = (r_n)'$ and $-x = (s_n)'$ with $r_n \geq 0$ and $s_n \geq 0$ for all n , then $(r_n + s_n)$ is a null sequence (because $x + (-x) = 0$); but $0 \leq r_n \leq r_n + s_n$, so (r_n) is also a null sequence, whence $x = 0$.}

(3) If $x \in \mathbb{R}$ then either $x \geq 0$ or $-x \geq 0$. {Proof: Suppose $x = (r_n)'$. If $r_n \geq 0$ ultimately then $x \geq 0$, and if $r_n \leq 0$ ultimately then $-x \geq 0$. It remains to consider the case that $r_n < 0$ frequently (say for the indices $n_1 < n_2 < n_3 < \dots$) and $r_n > 0$ frequently (say for the indices $m_1 < m_2 < m_3 < \dots$); we shall show that this implies $x = 0$, that is, (r_n) is null. Given any rational $t > 0$, choose an index N such that $m, n \geq N \Rightarrow |r_m - r_n| \leq t$. Choose k so that $m_k \geq N$ and $n_k \geq N$; then $|r_{m_k} - r_{n_k}| \leq t$, thus

$$0 < r_{m_k} < r_{m_k} - r_{n_k} = |r_{m_k} - r_{n_k}| \leq t,$$

and it follows that

$$n \geq N \Rightarrow |r_n| \leq |r_n - r_{m_k}| + |r_{m_k}| \leq 2t$$

whence the nullity of (r_n) .}

1.8.18. *Definition.* For $x, y \in \mathbb{R}$, write $x \leq y$ in case $y - x \geq 0$ in the sense of 1.8.16.

1.8.19. *Lemma.* For the relation \leq just defined, \mathbb{R} is an ordered field (in particular, \leq is a simple ordering of \mathbb{R}).

Proof. Immediate from 1.8.17.⁶ \diamond

1.8.20. *Remark.* For $r \in \mathbb{Q}$, $r \geq 0$ in $\mathbb{Q} \Leftrightarrow r' \geq 0$ in \mathbb{R} . {Proof: The implication \Rightarrow is trivial. Assuming $r' \geq 0$ let us show that $r \geq 0$. By assumption, $r' = (r_n)'$ with $r_n \geq 0$ for all n . Suppose to the contrary that $r < 0$. Since $(r - r_n)$ is null, $r_n - r < -r$ ultimately, whence $(r_n)' \leq 0$, that is, $r' \leq 0$; already $r' \geq 0$, so $r' = 0$, a contradiction.}

1.8.21. *Lemma.* If $x \in \mathbb{R}$, $x > 0$, then there exists a rational $r > 0$ such that $0 < r' < x$.

Proof. By assumption, $x = (r_n)'$ with $r_n \geq 0$ for all n , and (r_n) not a null sequence. Suppose to the contrary that no such r exists. Let r be

⁶ First course, Chapter 1, §2.

any rational number > 0 . By supposition, $r' \geq x$, so $r' - x = (s_n)'$ with $s_n \geq 0$ for all n . Then $(r_n + s_n - r)$ is null, therefore $r_n + s_n - r \leq r$ ultimately, whence $0 \leq r_n \leq r_n + s_n \leq 2r$ ultimately. To summarize, $(\forall r \in \mathbb{Q}, r > 0) 0 \leq r_n \leq 2r$ ultimately. This shows that (r_n) is null, a contradiction. \diamond

1.8.22. Since \mathbb{R} is an ordered field (1.8.19) we can define absolute values in \mathbb{R} :

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0. \end{cases}$$

{In view of 1.8.20, this definition is consistent with the earlier definition of absolute value for rationals, and $|r'| = |r|'$ for all $r \in \mathbb{Q}$.} The proof of 1.8.2 applies equally well for \mathbb{R} .

1.8.23. *Definition.* A sequence (x_k) in \mathbb{R} is said to **converge** to the **limit** $x \in \mathbb{R}$ if

$$(\forall \epsilon \in \mathbb{R}, \epsilon > 0) |x_k - x| \leq \epsilon \text{ ultimately.}$$

{In view of 1.8.21, it suffices to consider $\epsilon = r'$ with $r > 0$ rational.}

Such limits x are unique by the usual elementary argument, and one writes $x_k \rightarrow x$ (as $k \rightarrow \infty$).

1.8.24. *Lemma.* If $x_k \rightarrow x$ in \mathbb{R} and $x_k \geq 0$ for all k , then $x \geq 0$.

Proof. From the inequality $||x_k| - |x|| \leq |x_k - x|$ we see that $|x_k| \rightarrow |x|$; but $|x_k| = x_k \rightarrow x$, therefore $x = |x| \geq 0$ by the uniqueness of limits. \diamond

1.8.25. *Lemma.* If $x = (r_n)' \in \mathbb{R}$ then $r'_k \rightarrow x$ as $k \rightarrow \infty$.

Proof. For every rational $t > 0$, there is an index N such that $|r_m - r_n| \leq t$ for all $m, n \geq N$. For each $k \geq N$, $r'_k - x = (r_k - r_n)$ and $|r_k - r_n| \leq t$ for all $n \geq N$, therefore $|r'_k - x| \leq t'$. Since ϵ 's of the form t' suffice in the criterion for convergence (1.8.23), this shows that $r'_k \rightarrow x$. \diamond

1.8.26. *Theorem.* (Cauchy's criterion) Every Cauchy sequence in \mathbb{R} converges to an element of \mathbb{R} .

Proof. Let (x_k) be a Cauchy sequence in \mathbb{R} (for the definition of 'Cauchy sequence', paraphrase 1.8.7 with \mathbb{Q} replaced by \mathbb{R}). For each index k , choose $r_k \in \mathbb{Q}$ with $|r'_k - x_k| \leq 1/k'$ (1.8.25). Let $t \in \mathbb{Q}$, $t > 0$. Since (x_k) is Cauchy, there is an index N such that

$$(\forall j, k \geq N) |x_j - x_k| \leq t';$$

we can suppose further that $1/N \leq t$ (1.8.10 and 1.8.20). Then, for all $j, k \geq N$,

$$|r'_j - r'_k| \leq |r'_j - x_j| + |x_j - x_k| + |x_k - r'_k| \leq 1/j' + t' + 1/k' \leq 3t',$$

and, since $|r_j - r_k|' = |r_j' - r_k'|$, it follows from 1.8.20 that $|r_j - r_k| \leq 3t$. This shows that (r_n) is Cauchy in \mathbb{Q} . Let $x = (r_n)' = (r_n) + \mathcal{N}$; since

$$|x - x_k| \leq |x - r_k'| + |r_k' - x_k| \leq |x - r_k'| + 1/k'$$

and $r_k' \rightarrow x$ (1.8.25), it follows that $x_k \rightarrow x$. \diamond

There is no further need to distinguish between a rational number r and its image r' in \mathbb{R} ; henceforth we identify r with r' , regard \mathbb{Q} as a subfield of \mathbb{R} , and call the elements of \mathbb{R} **real numbers**. If $a, b \in \mathbb{R}$, $a \leq b$, we write $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ (called the **closed interval** with endpoints a and b). An easy consequence of Cauchy's criterion is the 'theorem on nested intervals':

1.8.27. Theorem. (Theorem on nested intervals) Suppose $[a_n, b_n]$ is a sequence of closed intervals in \mathbb{R} such that

$$[a_1, b_1] \supset [a_2, b_2] \supset [a_3, b_3] \supset \dots$$

and such that $b_n - a_n \rightarrow 0$. Then

- (i) $\bigcap [a_n, b_n] = \{c\}$ for some $c \in \mathbb{R}$, and
- (ii) if $x_n \in [a_n, b_n]$ for all n , then $x_n \rightarrow c$.

Proof. Suppose $x_n \in [a_n, b_n]$ for all n . If $m, n \geq N$ then x_m, x_n both belong to $[a_N, b_N]$, therefore $|x_m - x_n| \leq b_N - a_N$; since $b_n - a_n \rightarrow 0$, it follows that (x_n) is Cauchy, hence convergent (1.8.26), say $x_n \rightarrow x$. If also $y_n \in [a_n, b_n]$ for all n and $y_n \rightarrow y$, then $|x_n - y_n| \leq b_n - a_n$ shows that $x_n - y_n \rightarrow 0$, whence $x = y$. This proves both (i) and (ii). \diamond

1.8.28. Definition. A **Dedekind cut** of \mathbb{R} is a pair (A, B) of nonempty subsets of \mathbb{R} , with $A \cup B = \mathbb{R}$, such that $a < b$ for all $a \in A$ and $b \in B$.

It follows from the theorem on nested intervals that every Dedekind cut of \mathbb{R} is effected by an element of \mathbb{R} :

1.8.29. Theorem. If (A, B) is a Dedekind cut of \mathbb{R} , then there exists an element $c \in \mathbb{R}$ such that either

$$A = \{x \in \mathbb{R} : x \leq c\}, \quad B = \{x \in \mathbb{R} : x > c\}$$

or

$$A = \{x \in \mathbb{R} : x < c\}, \quad B = \{x \in \mathbb{R} : x \geq c\}.$$

Proof. The crux of the matter is to show that either A has a largest element or B has a smallest element. Choose $a_1 \in A$, $b_1 \in B$; let us construct a closed subinterval $[a_2, b_2] \subset [a_1, b_1]$ such that $a_2 \in A$, $b_2 \in B$ and $b_2 - a_2 = \frac{1}{2}(b_1 - a_1)$. Let $c_1 = \frac{1}{2}(a_1 + b_1)$. If $c_1 \in A$, let $a_2 = c_1$ and $b_2 = b_1$; the only alternative is that $c_1 \in B$, in which case we take $a_2 = a_1$ and $b_2 = c_1$. Repeating the process, we get a sequence

of closed intervals

$$[a_1, b_1] \supset [a_2, b_2] \supset [a_3, b_3] \supset \dots$$

such that $a_n \in A$, $b_n \in B$ and $b_n - a_n \rightarrow 0$. Let c be the unique real number such that $\bigcap [a_n, b_n] = \{c\}$ (1.8.27). In particular, $a_n \rightarrow c$ and $b_n \rightarrow c$.

case 1: $c \in A$

If $a \in A$ then $a < b_n$ for all n , therefore $a \leq c$ (1.8.24); thus $A \subset \{x \in \mathbb{R} : x \leq c\}$. On the other hand, if $b \in B$ then $c < b$ (because $c \in A$), thus $B \subset \{x \in \mathbb{R} : x > c\}$. It then follows from the law of trichotomy for \mathbb{R} (cf. 1.7.5) that both inclusions are equalities.

case 2: $c \in B$

One argues similarly that $B = \{x \in \mathbb{R} : x \geq c\}$ and $A = \{x \in \mathbb{R} : x < c\}$. \diamond

The end of our quest:

1.8.30. Theorem. \mathbb{R} is a complete ordered field.

Proof. Assuming S is a nonempty subset of \mathbb{R} that is bounded above, we must show that the (nonempty) set B of majorants of S has a smallest element.

If $b \in B$ then B contains every $x \in \mathbb{R}$ with $x > b$. On the other hand, $B \neq \mathbb{R}$; indeed, if $x \in S$ then $x - 1 \notin B$. Let $A = \mathbb{R} - B$ (nonempty, by the preceding remark). If $a \in A$ and $b \in B$ then $a < b$ (the alternative, $a \geq b \in B$, would place a in B). Thus (A, B) is a Dedekind cut of \mathbb{R} .

Let c be the unique element of \mathbb{R} provided by 1.8.29; it will suffice to show that $c \in B$. Assume to the contrary that $c \in A$, that is, c does not majorize S ; then S contains an element x with $x > c$. The element $y = \frac{1}{2}(x + c)$ belongs to B (because $y > c$) hence majorizes S ; in particular, $y \geq x$, which contradicts the definition of y . \diamond

1.8.31. *Terminology.* A function whose range consists of real numbers is said to be **real-valued**; a function whose domain consists of real numbers is said to be a function of a **real variable**. For example, the function $f : (0, 1] \rightarrow \mathbb{R}$ defined by $f(x) = 1/x$ is a real-valued function of a real variable.

Exercises

1. Show that the ordering of \mathbb{R} is Archimedean: if $x, y \in \mathbb{R}$ and $x > 0$, then there exists a positive integer n such that $nx > y$.

{Hint: 1.8.21.}

2. Let $x, y \in \mathbb{R}$ and let n be a positive integer. If $x \geq 0$ and $y \geq 0$, then

- (i) $x < y \Leftrightarrow x^n < y^n$,
- (ii) $x = y \Leftrightarrow x^n = y^n$,
- (iii) $x > y \Leftrightarrow x^n > y^n$.

If n is odd, then the restriction that x and y are nonnegative can be dropped.

3. If a, b, c are nonnegative real numbers such that $a \leq b + c$, then

$$\frac{a}{1+a} \leq \frac{b}{1+b} + \frac{c}{1+c}.$$

1.9. Finite and Infinite Sets

One approach is to call a set E 'finite' if either $E = \emptyset$ or there exists a surjection $\{1, 2, \dots, n\} \rightarrow E$ for some positive integer n , then define the 'infinite' sets to be those that are not finite.¹ In the following 'dual' approach,² finite and infinite sets are defined without reference to the set of positive integers (but, of course, must eventually make contact with it). The main objective is to rehearse some of the most frequently used arguments in discussions involving finiteness and infiniteness.

1.9.1. *Definition.* (R. Dedekind) A set E is said to be **infinite** if there exists an injection $E \rightarrow E$ that is not surjective. A set that is not infinite is said to be **finite**.

1.9.2. *Examples.* (1) The set \mathbb{P} of all positive integers is infinite (consider the mapping $n \mapsto n + 1$).

(2) The empty set is finite 'by default' (for, by the convention in 1.3.1, the empty set is not the domain of a function). For $E \neq \emptyset$, the following conditions are equivalent: (a) E is finite; (b) every injection $E \rightarrow E$ is surjective.

(3) Singletons $\{c\}$ are finite ($c \mapsto c$ is the only available mapping).

1.9.3. *Theorem.* (1) Every superset of an infinite set is infinite.

(1') Every subset of a finite set is finite.

Let $f : E \rightarrow F$.

(2) If f is bijective and E is infinite, then F is infinite.

(2') If f is bijective and E is finite, then F is finite.

(3) If f is injective and E is infinite, then F is infinite.

(3') If f is injective and F is finite, then E is finite.

(4) If f is surjective and E is finite, then F is finite.

(4') If f is surjective and F is infinite, then E is infinite.

¹ First course, §4.4, or E. Hewitt and K. Stromberg, *Real and abstract analysis* [Springer-Verlag, New York, 1965], p. 21, (4.12).

² See A. Abian, *The theory of sets and transfinite arithmetic* [Saunders, Philadelphia, 1965].

Proof. (1) Suppose E is infinite and $F \supset E$. If $F = E$ there is nothing to prove; assume $F \neq E$. By assumption, there exists an injection $f : E \rightarrow E$ that is not surjective. The function $f' : F \rightarrow F$ defined by

$$f'(x) = \begin{cases} f(x) & \text{for } x \in E \\ x & \text{for } x \in F - E. \end{cases}$$

is injective but not surjective.

(1') This is (1) in contrapositive form.

(2) Suppose E is infinite and $f : E \rightarrow F$ is bijective. Let $g : E \rightarrow E$ be an injection that is not surjective. The mapping $f \circ g \circ f^{-1} : F \rightarrow F$ is injective (it is a composite of injectives) but not surjective (if it were, then $f^{-1} \circ (f \circ g \circ f^{-1}) \circ f = g$ would be surjective).

(2') If $f : E \rightarrow F$ is bijective, then so is $f^{-1} : F \rightarrow E$, thus (2') is the contrapositive form of (2).

(3) If E is infinite and $f : E \rightarrow F$ is injective, then f defines a bijection $E \rightarrow f(E)$, therefore $f(E)$ is infinite by (2); then $F \supset f(E)$ is infinite by (1).

(3') The contrapositive form of (3).

(4), (4') Suppose $f : E \rightarrow F$ is surjective. By 1.5.4 (the axiom of choice plays a role here), there exists an injection $g : F \rightarrow E$, thus

$$E \text{ finite} \Rightarrow F \text{ finite}$$

by (3'), and F infinite $\Rightarrow E$ infinite by (3). \diamond

1.9.4. Theorem. *A set E is infinite if and only if there exists an injection $\mathbb{P} \rightarrow E$.*

Proof. "If": If there exists an injection $f : \mathbb{P} \rightarrow E$ then, since \mathbb{P} is infinite (1.9.2), so is E by (3) of the preceding theorem.

"Only if": Assuming E is infinite, let $f : E \rightarrow E$ be an injection that is not surjective and choose a point $z \in E - f(E)$. Define $g : \mathbb{P} \rightarrow E$ by the formula $g(n) = f^n(z)$, where the 'composition powers' f^n are defined recursively by $f^1 = f$ and $f^{n+1} = f^n \circ f$. We assert that g is injective. Assuming $m, n \in \mathbb{P}$, $m < n$, we must show that $g(m) \neq g(n)$. Let $p = n - m$; then

$$g(n) = f^{m+p}(z) = f^m(f^p(z)),$$

whereas $g(m) = f^m(z)$; assuming to the contrary that $g(m) = g(n)$, that is, $f^m(z) = f^m(f^p(z))$, it follows from the injectivity of f that $z = f^p(z) \in f(E)$, contrary to the choice of z . \diamond

For each positive integer n , we write $\mathbb{P}_n = \{1, \dots, n\}$ (the set of all integers from 1 to n). The next target (1.9.13): A nonempty set E is finite if and only if it is bijective with some \mathbb{P}_n .

1.9.5. Lemma. *If S is a nonempty subset of \mathbb{P} with no largest element, then S is infinite.*

Proof. For each $a \in S$ write $S(a) = \{k \in S : k > a\}$; by assumption, $S(a) \neq \emptyset$. Define $f : S \rightarrow S$ as follows: for each $a \in S$, let $f(a)$ be the smallest element of $S(a)$. In particular, $f(a) > a$ for all $a \in S$. We show that S is infinite by verifying that f is injective but not surjective.

f is injective: For, if $a, b \in S$ and $a < b$, then $b \in S(a)$, therefore $f(a) \leq b < f(b)$.

f is not surjective: For, if z is the smallest element of S , then $z \leq a < f(a)$ for all $a \in S$, therefore $z \notin f(S)$. \diamond

1.9.6. Lemma. *If A is a finite subset of \mathbb{P} , then $\mathbb{C}A = \mathbb{P} - A$ has no largest element.*

Proof. Write $B = \mathbb{C}A$. Arguing contrapositively, let us show that if B has a largest element m , then $\mathbb{C}B$ is infinite. By assumption, $B \subset \{1, \dots, m\} = \mathbb{P}_m$, therefore $\mathbb{C}B \supset \mathbb{C}\mathbb{P}_m = \{k \in \mathbb{P} : k > m\}$; since $\mathbb{C}\mathbb{P}_m$ is infinite (consider the map $k \mapsto k + 1$), so is its superset $\mathbb{C}B$ (1.9.3). \diamond

1.9.7. Lemma. *If A is a finite subset of \mathbb{P} , then $\mathbb{C}A$ is infinite.*

Proof. By the preceding lemma, $\mathbb{C}A$ has no largest element, therefore is infinite (1.9.5). \diamond

1.9.8. Lemma. *\mathbb{P} is not the union of two finite sets.*

Proof. Assuming $\mathbb{P} = A \cup B$ with A finite, we must show that B is infinite. We have $B \supset \mathbb{P} - A$ and $\mathbb{P} - A$ is infinite (1.9.7), therefore so is its superset B . \diamond

1.9.9. Lemma. *If $f : E \rightarrow F$ and A is a finite subset of E , then $f(A)$ is finite.*

Proof. The restriction of f to A defines a surjection $A \rightarrow f(A)$, therefore $f(A)$ is finite by (4) of 1.9.3. \diamond

1.9.10. Lemma. *If A and B are finite sets, then $A \cup B$ is finite.*

Proof. Let $E = A \cup B$ (cf. 1.4.1) and assume to the contrary that E is infinite. By 1.9.4, there exists an injection $f : \mathbb{P} \rightarrow E$, therefore a surjection $g : E \rightarrow \mathbb{P}$ (1.5.4); then $\mathbb{P} = g(E) = g(A) \cup g(B)$ is the union of two finite sets (1.9.9), contrary to 1.9.8. \diamond

1.9.11. Lemma. *For every positive integer n , \mathbb{P}_n is finite.**

Proof. (by induction on n). $\mathbb{P}_1 = \{1\}$ and $\mathbb{P}_{n+1} = \mathbb{P}_n \cup \{n + 1\}$; cite 1.9.2, (3) and the preceding lemma. \diamond

1.9.12. Lemma. *Let $m, n \in \mathbb{P}$. If there exists a bijection $\mathbb{P}_n \rightarrow \mathbb{P}_m$, then $m = n$.*

Proof. Let $f : \mathbb{P}_n \rightarrow \mathbb{P}_m$ be a bijection. We can suppose that $n \geq m$ (if $m \geq n$, consider instead the inverse bijection f^{-1}); then $\mathbb{P}_n \supset \mathbb{P}_m$. Consider the mappings

$$\mathbb{P}_n \xrightarrow{f} \mathbb{P}_m \xrightarrow{i} \mathbb{P}_n,$$

where i is the insertion mapping. The composite function $i \circ f : \mathbb{P}_n \rightarrow \mathbb{P}_n$ is injective (because f and i are), therefore bijective (1.9.11); then $(i \circ f) \circ f^{-1} = i$ is also bijective, therefore $\mathbb{P}_m = \mathbb{P}_n$ and $m = n$ (k is the largest element of \mathbb{P}_k). \diamond

1.9.13. Theorem. *A nonempty set is finite if and only if it is bijective with some \mathbb{P}_n .*

Proof. “If”: Suppose $f : E \rightarrow \mathbb{P}_n$ is bijective; since \mathbb{P}_n is finite (1.9.11), so is E (1.9.3).

“Only if”: Arguing contrapositively, if E is a nonempty set that is not bijective with any \mathbb{P}_n , we must show that E is infinite. By 1.9.4, it suffices to find an injection $\mathbb{P} \rightarrow E$, in other words, a sequence (x_n) in E such that $n \mapsto x_n$ is injective (that is, the x_n are pairwise distinct).

An informal, ‘recursive’ argument for producing such a sequence is as follows. Choose $x_1 \in E$ and let $E_1 = \{x_1\}$. Let $n \geq 1$ and assume already chosen distinct points x_1, \dots, x_n of E . Let $E_n = \{x_1, \dots, x_n\}$. Then $i \mapsto x_i$ is a bijection $\mathbb{P}_n \rightarrow E_n$, so by hypothesis $E_n \neq E$; choose $x_{n+1} \in E - E_n$. “And so on ...”³ \diamond

1.9.14. Definition. If a set E is bijective with some \mathbb{P}_n then n is unique (1.9.12); we then say that E has **cardinality** n and we write $\text{card } E = n$. Convention: $\text{card } \emptyset = 0$.

In particular, $\text{card } \mathbb{P}_n = n$ for all $n \in \mathbb{P}$.

The use of the symbol $\text{card } E$ in connection with infinite sets E is discussed in §1.12. In the next section, we study the infinite set \mathbb{P} in greater detail.

Exercises

1. For a nonempty set E , the following conditions are equivalent: (a) E is finite; (b) every surjection $E \rightarrow E$ is injective.

2. Let E_1, \dots, E_r be nonempty sets and let $E = E_1 \times \dots \times E_r$. Show that E is finite if and only if every E_k is finite.

³For a formal (i.e., honest) argument, see Hewitt and Stromberg [*op. cit.*], p. 22, (4.15), or P. R. Halmos, *Naïve set theory* [Springer-Verlag, New York, 1974], pp. 60-61.

3. (i) Every nonempty finite partially ordered set has at least one minimal element (1.7.10).

(ii) A finite pre-ordered set need not have a minimal element.

4. As an alternative to Definition 1.9.1, call a set E *infinite* if there exists an injection $\mathbb{P} \rightarrow E$, and *finite* if either $E = \emptyset$ or there exists a surjection $\{1, \dots, n\} \rightarrow E$ for some $n \in \mathbb{P}$. Work out a proof that a set is infinite if and only if it is not finite.

1.10. Countable and Uncountable Sets

1.10.1. *Definition.* A set E is said to be **countable** if either $E = \emptyset$ or there exists a surjection $\mathbb{P} \rightarrow E$. {In other words, either $E = \emptyset$ or there exists a sequence (x_n) in E such that every point of E is equal to at least one term of the sequence.} If E is not countable it is said to be **uncountable**. (We shall see in 1.10.11 that the field \mathbb{R} of real numbers is uncountable.)

1.10.2. *Remarks.* 1. If E is countable (uncountable) and if $f : E \rightarrow F$ is bijective, then F is also countable (uncountable).

2. Every subset of a countable set is countable (hence every superset of an uncountable set is uncountable).

{Proof: Suppose $f : \mathbb{P} \rightarrow E$ is surjective and A is a nonempty subset of E ; we have to find a surjection $\mathbb{P} \rightarrow A$. Choose a point $a \in A$ and let $g : \mathbb{P} \rightarrow A$ be the function such that $g(n) = f(n)$ if $f(n) \in A$, and $g(n) = a$ if $f(n) \notin A$. For every $x \in A$, there exists an n with $f(n) = x$, therefore $x = g(n)$; thus g is surjective.}

3. Every finite set is countable. {A nonempty finite set is bijective with some \mathbb{P}_n (1.9.13), hence is countable by Remarks 1 and 2.}

On the way to characterizing the countable sets:

1.10.3. *Lemma.* Every infinite subset of \mathbb{P} is bijective with \mathbb{P} .

Proof. Assuming $A \subset \mathbb{P}$ infinite, we must construct a bijection $f : \mathbb{P} \rightarrow A$. Define $f(1)$ to be the smallest element of A and, recursively, $f(n+1)$ to be the smallest element of $A - \{f(1), \dots, f(n)\}$. It is obvious that f is injective. Better yet, it is strictly increasing. {Proof: $f(2) > f(1)$ because $f(1)$ is the smallest element of A and $f(2) \neq f(1)$; $f(3) > f(2)$ because there are no elements of A between $f(1)$ and $f(2)$, and none less than $f(1)$. In general, there are no elements of A between $f(i)$ and $f(i+1)$ for any i , whence $f(n+1) > f(n)$.} An easy induction then shows that $f(n) \geq n$ for every n .

To complete the proof, we need only show that f is surjective. Assume to the contrary that $A - f(\mathbb{P})$ contains some element k . In particular,

$k \in A - \{f(1), \dots, f(k)\}$, so $k \geq f(k+1)$ by the minimality of $f(k+1)$, whence the absurdity $k+1 \leq f(k+1) \leq k$. \diamond

1.10.4. Theorem. *A set E is countable if and only if either (1) E is bijective with \mathbb{P} , or (2) E is bijective with $\mathbb{P}_n = \{1, \dots, n\}$ for some positive integer n , or (3) $E = \emptyset$.*

Proof. “If”: That a set satisfying (1) or (3) is countable is obvious from the definition, and a set satisfying (2) is countable by Remarks 1 and 2 of 1.10.2.

“Only if”: Suppose $E \neq \emptyset$ and $f: \mathbb{P} \rightarrow E$ is surjective; we have to show that E satisfies either (1) or (2). If E is finite, we are done (1.9.13). Suppose E is infinite. For every $x \in E$ let $A_x = f^{-1}(\{x\})$ (a nonempty subset of \mathbb{P} , since f is surjective) and let $g(x)$ be the smallest element of A_x ; this defines a function $g: E \rightarrow \mathbb{P}$, injective since $A_x \cap A_y = \emptyset$ when $x \neq y$. Since g is injective, E is bijective with $g(E)$; since E is infinite, so is $g(E)$; thus $g(E)$ is bijective with \mathbb{P} by the lemma, therefore so is E . \diamond

1.10.5. Definition. In view of 1.10.4, a set that is bijective with \mathbb{P} is said to be **countably infinite** (or **denumerably infinite**, or, simply, **denumerable**).

The rest of the section is devoted to examples of countable and uncountable sets. The following result is a ready source of uncountable sets:

1.10.6. Theorem. *If E is a set and $\mathcal{P}(E)$ is its power set (1.4.1), then there does not exist a surjective mapping $E \rightarrow \mathcal{P}(E)$.*

Proof. {For finite sets, this is not news: If E has n elements then $\mathcal{P}(E)$ has 2^n elements, and $n < 2^n$. If $E = \emptyset$ then $\mathcal{P}(E) = \{\emptyset\}$ is nonempty, so even if one admitted the empty mapping $\emptyset \rightarrow \mathcal{P}(\emptyset)$, it would not be surjective.} Assume to the contrary that there exists a set E that admits a surjective mapping $f: E \rightarrow \mathcal{P}(E)$. Let

$$A = \{x \in E: x \notin f(x)\}.$$

Since f is surjective, $A = f(a)$ for some $a \in E$. Either (i) $a \in A$ or (ii) $a \notin A$, but both alternatives lead to a contradiction: (i) if $a \in A = f(a)$ then $a \notin A$ by the definition of A ; (ii) if $a \notin A = f(a)$, then $a \in A$ by the definition of A . \diamond

1.10.7. Corollary. $\mathcal{P}(\mathbb{P})$ is uncountable.

Proof. By the theorem, there exists no surjection $\mathbb{P} \rightarrow \mathcal{P}(\mathbb{P})$. \diamond

1.10.8. Theorem. $\mathbb{P} \times \mathbb{P}$ is bijective with \mathbb{P} .

Proof. The mapping $f : \mathbb{P} \times \mathbb{P} \rightarrow \mathbb{P}$ defined by $f(m, n) = 2^m 3^n$ is injective, so $\mathbb{P} \times \mathbb{P}$ is bijective with its range $A = f(\mathbb{P} \times \mathbb{P})$; but A is infinite (the mapping $m \mapsto 2^m$ is an injection $\mathbb{P} \rightarrow A$), therefore A is bijective with \mathbb{P} (1.10.3). \diamond

1.10.9. Corollary. *If E and F are countable sets then so is $E \times F$.*

Proof. If $E = \emptyset$ or $F = \emptyset$ then $E \times F = \emptyset$. Assuming E and F nonempty, by hypothesis there exist surjections $g : \mathbb{P} \rightarrow E$ and $h : \mathbb{P} \rightarrow F$; the mapping $\mathbb{P} \times \mathbb{P} \rightarrow E \times F$ defined by $(m, n) \mapsto (g(m), h(n))$ is surjective, and by the theorem there exists a surjection $\mathbb{P} \rightarrow \mathbb{P} \times \mathbb{P}$, so composition yields a surjection $\mathbb{P} \rightarrow E \times F$. \diamond

By an obvious induction, every finite product $E_1 \times \dots \times E_n$ of countable sets is countable.

1.10.10. Corollary. *The field \mathbb{Q} of rational numbers is denumerable.*

Proof. If $S = \{m/n : m, n \in \mathbb{P}\}$ (the set of all positive rational numbers), then $\mathbb{P} \subset S$ and the mapping $(m, n) \mapsto m/n$ is a surjection $\mathbb{P} \times \mathbb{P} \rightarrow S$, thus S is infinite and countable (1.10.8), therefore denumerable (1.10.3). Let $f : \mathbb{P} \rightarrow S$ be a bijection and write $r_n = f(n)$ ($n \in \mathbb{P}$). Then the sequence

$$0, r_1, -r_1, r_2, -r_2, r_3, -r_3, \dots$$

exhausts \mathbb{Q} and defines a bijection $\mathbb{P} \rightarrow \mathbb{Q}$. \diamond

1.10.11. Corollary. *The field \mathbb{R} of real numbers is uncountable.*

Proof. It suffices to show that $[0, 1]$ is uncountable (1.10.2). For this, we need only show that every mapping $f : \mathbb{P} \rightarrow [0, 1]$ fails to be surjective. For any closed interval $[a, b]$, let c and d be its points of trisection ($a < c < d < b$) and call $[a, c]$, $[c, d]$, $[d, b]$ the *closed thirds* (left, middle and right, respectively) of $[a, b]$. Define recursively a decreasing sequence $I_1 \supset I_2 \supset I_3 \supset \dots$ of nondegenerate closed subintervals of $[0, 1]$ as follows: I_1 is a closed third of $[0, 1]$ such that $f(1) \notin I_1$ (to be definite, we could choose I_1 to be the left-most third with this property), and, recursively, I_{n+1} is a closed third of I_n such that $f(n+1) \notin I_{n+1}$. By the theorem on nested intervals (1.8.27), the intersection $\bigcap I_n$ contains a point x (in fact, only one). For every n , we have $x \in I_n$, therefore $f(n) \neq x$, thus x fails to belong to the range of f . \diamond

Exercises.

1. A finite product of denumerable sets is denumerable.
2. For every infinite set E , $\mathcal{P}(E)$ is uncountable.
{Hint: 1.9.4 and 1.10.7.}
3. The union of a sequence of countable sets is countable.

1.11. Zorn's Lemma, the Well-Ordering Theorem

The Axiom of Choice (AC), introduced in the context of product sets (§1.4), already figures in our discussion of the algebra of functions (1.5.1, 1.5.2) and the mapping properties of finiteness and infiniteness (see (4) and (4') of 1.9.3).

In this section we record several other statements, equivalent to the Axiom of Choice, that have many applications in algebra, topology and analysis. Ideally, such matters are taken up in a leisurely preliminary course¹; realistically, the reader who has had such a course is probably not reading this chapter. A detour into the foundations of mathematics (the homeland of the axioms of set theory) would run the risk of turning into a derailment; the present section is a skeletal compromise: the equivalent forms of the Axiom of Choice are stated for reference, omitting the proofs that they are equivalent but including some sample applications. First, a little motivation.

One equivalent axiom (AC') already mentioned asserts the existence of 'choice functions' (1.4.6); here is a situation where no axiom is needed because a theorem is available:

1.11.1. Theorem. *Every nonempty countable set admits a choice function.*

Proof. Assuming $f : \mathbb{P} \rightarrow E$ surjective (1.10.1) we seek a function $c : \mathcal{P}(E) - \{\emptyset\} \rightarrow E$ such that $c(A) \in A$ for every nonempty subset A of E . For each nonempty $A \subset E$, $f^{-1}(A)$ is a nonempty subset of \mathbb{P} ; by the principle of mathematical induction (1.7.12), $f^{-1}(A)$ has a smallest element $\min f^{-1}(A)$, and the formula $c(A) = f(\min f^{-1}(A))$ defines a choice function for E . {The structure of the proof is that \mathbb{P} has a choice function (namely $S \mapsto \min S$), therefore so does every functional image of \mathbb{P} .} \diamond

The property of the ordered set \mathbb{P} that makes the foregoing argument work is distilled in the following definition:

1.11.2. *Definition.* A partially ordered set (X, \leq) is said to be **well-ordered** if every nonempty subset of X has a smallest (or "first") element; that is, if $\emptyset \neq S \subset X$ then there exists an element $a \in S$ such that $a \leq x$ for all $x \in S$. Such a relation \leq is called a **well-ordering** of X .

1.11.3. If E is a nonempty set that can be indexed (1.3.8) by a well-ordered set, then E admits a choice function (cf. the proof of 1.11.1).

The following property of well-ordered sets is the basis for inductive arguments in such sets:

¹ Based, for example, on I. Kaplansky's *Set theory and metric spaces* [Chelsea, New York, 1977] or the first two chapters of E. Hewitt and K. Stromberg's *Real and abstract analysis* [Springer-Verlag, New York, 1965].

1.11.4. Theorem. (Principle of transfinite induction) *Let E be a well-ordered set. If S is a subset of E satisfying the condition*

$$(*) \quad (y < x \Rightarrow y \in S) \Rightarrow x \in S$$

then $S = E$.

Proof. The meaning of $(*)$ in words: if x is an element of E such that S contains every element of E that is smaller than x , then S must also contain x . {There is no harm if $S = \emptyset$; this simply forces $E = \emptyset$.}

Arguing contrapositively, it is the same to show that a proper subset of E cannot have the property $(*)$. Let S be a proper subset of E and let x be the smallest element of $E - S$. If $y < x$ then $y \notin E - S$ by the minimality of x , thus $y \in S$; S contains every element smaller than x but it does not contain x , so S does not have the property $(*)$. \diamond

The most accessible applications of transfinite induction are to the theory of well-ordered sets itself; numerous examples are given in §1.13.

Georg Cantor (1845-1918), the founder of the theory of sets, conjectured (in 1883) that every set can be equipped with a well-ordering. This conjecture,

(WO) Every set can be well ordered

is known as the **well-ordering theorem**. In 1904, Ernst Zermelo (1871-1953) proved that if the Axiom of Choice is admitted then Cantor's conjecture is true. Zermelo's theorem is the implication $(AC) \Rightarrow (WO)$. The reverse implication is trivial (1.11.3), thus

1.11.5. (Zermelo's) Theorem. $(AC) \Leftrightarrow (WO)$.

An equivalent axiom, frequently easier to apply than the Axiom of Choice or well-ordering, was proposed in 1935 by Max Zorn (1906-1993); this axiom, now known as **Zorn's Lemma**, is conveniently stated in terms of the following concept:

1.11.6. *Definition.* A partially ordered set (X, \leq) is said to be **inductive** if every simply ordered subset of X has an upper bound in X .

Zorn's lemma:

(ZL) A nonempty, inductive partially ordered set has at least one maximal element.

To summarize, $(AC) \Leftrightarrow (ZL) \Leftrightarrow (WO)$.

Two other equivalents of the Axiom of Choice are (H) **Hausdorff's maximality principle**, and (T) **Tukey's lemma**. Hausdorff's principle is the following proposition:

(H) Let (X, \leq) be any partially ordered set; let \mathcal{X} be the set of all simply ordered subsets of X , and order \mathcal{X} by inclusion. Then \mathcal{X} has a

maximal element. (Briefly, every partially ordered set contains a maximal chain.)

Tukey's lemma involves the following concept:

1.11.7. *Definition.* A set \mathcal{F} of sets is said to be of **finite character** if, for a set A ,

$$A \in \mathcal{F} \Leftrightarrow \text{every finite subset of } A \text{ belongs to } \mathcal{F}.$$

Tukey's lemma:

(T) Let \mathcal{F} be a nonempty set of sets, order \mathcal{F} by inclusion, and suppose that \mathcal{F} is of finite character. Then \mathcal{F} has a maximal element.

Packaging it all into one statement,

1.11.8. Theorem. $(AC) \Leftrightarrow (ZL) \Leftrightarrow (T) \Leftrightarrow (H) \Leftrightarrow (WO)$.

Variouslly called "axiom", "lemma", "principle", "theorem", each of these statements is in fact an axiom, consistent with and independent of the most widely accepted system of axioms for mathematics². The reader will find elegant and efficient proofs of these equivalences in the books of Kaplansky and Hewitt-Stromberg cited earlier. (The details are "elementary"—it is easy to follow the proofs step by step—but fiendishly ingenious.) The best way to get a feeling for the axioms is to work through some applications; we conclude the section with two such applications, the first to vector spaces (every vector space has a basis), the second for use later on in the theory of cardinality (given any two sets, one of them contains a copy of the other).

1.11.9. Theorem. *Every vector space has a basis.*

Proof. # 1 (assuming the well-ordering theorem). Let V be a vector space. A subset A of V is (linearly) *independent* if no element of A is a linear combination of the remaining elements of A , *generating* if every vector in V is a linear combination of elements of A , and a *basis* of V if it is both independent and generating.

If V contains only the zero vector θ then the empty set serves as basis. Assuming $V \neq \{\theta\}$, *well-order* the nonempty set $V - \{\theta\}$ and let

$$B = \{x \in V - \{\theta\} : x \text{ is not a linear combination of vectors } < x\};$$

we will show that (i) B is independent, and (ii) B is generating.

(i) Assuming to the contrary that B is not independent, there exists a linear relation $c_1x_1 + \dots + c_nx_n = \theta$ with x_1, \dots, x_n distinct elements of B and c_1, \dots, c_n nonzero scalars. If x_j is the *largest* of these vectors, then x_j is a linear combination of vectors $< x_j$, therefore $x_j \notin B$ (by the definition of B), a contradiction.

²I. Kaplansky, *op. cit.*, p. 59.

(ii) Assume to the contrary that some vector in V fails to be a linear combination of elements of B (hence is nonzero) and let x be the smallest such vector. In particular, $x \notin B$, therefore x is a linear combination $x = c_1x_1 + \dots + c_nx_n$ with $x_i < x$ for all i . By the minimality of x , every x_i is a linear combination of elements of B ; but then so is x , a contradiction. \diamond

Proof. # 2 (assuming Tukey's lemma). As in the preceding proof, we can suppose that V is a vector space $\neq \{\theta\}$. Let \mathcal{S} be the set of all independent subsets of V (for example, $\{x\} \in \mathcal{S}$ for every nonzero vector x). It is clear from the definition of independence that \mathcal{S} is of finite character, that is, a set $A \subset V$ is independent if and only if every finite subset of A is independent. By Tukey's lemma, \mathcal{S} has an element B that is maximal (for the inclusion relation). In particular, B is independent, so we need only show that it is generating. Indeed, if some vector $x \in V$ fails to be a linear combination of elements of B (so that, in particular, $x \notin B$), then $B \cup \{x\}$ is an independent set containing B properly, contrary to the maximality of B . \diamond

1.11.10. Before giving an application of Zorn's lemma, we will exhibit the inductive sets that enter into the proof. First, some notations are in order. Given a pair of sets E and F , we consider the set \mathcal{F} of all functions $u : A \rightarrow F$ defined on subsets of A of E . A nonempty subset G of $E \times F$ is the graph of such a function $u : A \rightarrow F$ if and only if it has the property

$$(x, y), (x, y') \in G \Rightarrow y = y',$$

in which case

$$A = \text{pr}_1 G = \{x : (x, y) \in G \text{ for some } y \in F\}$$

and, for $x \in A$, $u(x)$ is the unique element of F such that $(x, u(x)) \in G$. We write G_u for the graph of a function $u \in \mathcal{F}$, and \mathcal{G} for the set of all graphs obtained in this way:

$$\mathcal{G} = \{G_u : u \in \mathcal{F}\};$$

the mapping $u \mapsto G_u$ is a bijection $\mathcal{F} \rightarrow \mathcal{G}$. The set \mathcal{G} is partially ordered by inclusion (because $\mathcal{P}(E \times F)$ is); explicitly, the meaning of $G_u \subset G_v$ is that v is an *extension* of u , that is, the domain of u is contained in the domain of v (because $\text{pr}_1 G_u \subset \text{pr}_1 G_v$) and, for every x in the domain of u , $u(x) = v(x)$. We also consider the set

$$\mathcal{F}_{\text{inj}} = \{u \in \mathcal{F} : u \text{ is injective}\}$$

and write \mathcal{G}_{inj} for the corresponding set of graphs.

1.11.11. Lemma. *With the preceding notations, the partially ordered sets \mathcal{G} and \mathcal{G}_{inj} are inductive.*

Proof. We first prove that \mathcal{G} is inductive. Assuming \mathcal{C} is a chain in \mathcal{G} , we must show that \mathcal{C} has an upper bound in \mathcal{G} . Let $H = \bigcup \mathcal{C}$ be the union of all the sets $G \in \mathcal{C}$. Since $H \supset G$ for all $G \in \mathcal{C}$, it will suffice to show that H is the graph of a function. Suppose $(x, y), (x, y') \in H$, say

$$(x, y) \in G \in \mathcal{C} \text{ and } (x, y') \in G' \in \mathcal{C}.$$

Since \mathcal{C} is a chain, either $G \subset G'$ or $G' \subset G$; if, for example, $G \subset G'$, then (x, y) and (x, y') both belong to G' , therefore $y = y'$ because G' is the graph of a function.

Finally, assuming in addition that $\mathcal{C} \subset \mathcal{G}_{\text{inj}}$, we must show that H is the graph of an injective function. Suppose $(x, y), (x', y) \in H$; we are to show that $x = x'$. Say

$$(x, y) \in G \in \mathcal{C} \text{ and } (x', y) \in G' \in \mathcal{C};$$

if, for example, $G \subset G'$, then $x = x'$ because G' is the graph of an injective function. \diamond

1.11.12. Theorem. *If E and F are nonempty sets, then either there exists an injection $E \rightarrow F$ or there exists an injection $F \rightarrow E$.*

Proof (using Zorn's lemma). With notations as in the lemma, we know that \mathcal{G}_{inj} is inductive, so by Zorn's lemma it has a maximal element G . Say G is the graph of $u: A \rightarrow F$.

If $A = E$ we are done: u is an injection $E \rightarrow F$. Similarly, if $u(A) = F$ then $u(a) \mapsto a$ ($a \in A$) is an injection $F \rightarrow E$. Thus, we need only show that either $A = E$ or $u(A) = F$.

Assume to the contrary that $A \neq E$ and $u(A) \neq F$. Choose elements $x \in E - A$ and $y \in F - u(A)$; then $G \cup \{(x, y)\}$ is the graph of an injection $A \cup \{x\} \rightarrow F$, which contradicts the maximality of G . \diamond

Exercises

1. Every finite simply ordered set is well-ordered.

2. Let $(X_i)_{i \in I}$ be a family of partially ordered sets (1.7.1) and suppose that the index set I is well-ordered. Let $X = \prod X_i$ be the product set. For points $x = (x_i), y = (y_i)$ of X , define $x \leq y$ to mean that either (1) $x = y$, or (2) $x \neq y$ and $x_j < y_j$ for the smallest index j such that $x_j \neq y_j$ (cf. §1.7, Exercise 4).

(i) The relation \leq in X is a partial ordering (called the *lexicographic* ordering).

(ii) If the X_i are all simply ordered, then so is X .

3. If X and Y are well-ordered sets, then $X \times Y$ is well-ordered for the lexicographic ordering.

{Hint: If A is a nonempty subset of $X \times Y$, then the set $\text{pr}_1 A$ of first coordinates of the elements of A is nonempty. Let x_1 be the first element of $\text{pr}_1 A$, let B be the set of all elements of A whose first coordinate is x_1 , and let y_1 be the first element of $\text{pr}_2 B$. Argue that (x_1, y_1) is the first element of A .}

4. If X_1, \dots, X_n are well-ordered sets, then $X_1 \times \dots \times X_n$ is well-ordered for the lexicographic order.

{Hint: Exercise 3 and induction (cf. §1.7, Exercise 4).}

5. In Exercise 2, consider $I = \mathbb{P}$ and $X_n = \{1, 2\}$ ($n \in \mathbb{P}$), all with the usual order, and let $X = \prod X_n$ be the product set, equipped with the lexicographic order. If $e_n \in X$ is the element with 2 in the n 'th coordinate and 1's elsewhere, then $e_1 > e_2 > e_3 > \dots$, thus X is not well-ordered.

1.12. Cardinality

The concept of cardinality has been broached in connection with finite sets (1.9.14); our aim in this section is to reformulate the discussion in terms applicable to arbitrary sets. The discussion of cardinality for finite sets is expedited by the fact that we have at hand a full list of 'models' (up to a bijection) of such sets:

$$\emptyset, \{1\}, \{1, 2\}, \{1, 2, 3\}, \dots$$

Thus, when we observe that a finite set E is bijective with $\mathbb{P}_4 = \{1, 2, 3, 4\}$, we already have in hand a symbol 4 that we can declare to be 'the number of elements of E ' (or the 'cardinal number of E '). For infinite sets, we will have to make up such symbols as we go along. In the absence of symbols, the place to begin is with bijections.

1.12.1. *Definition.* For sets E and F , we write $E \sim F$ if there exists a bijection $E \rightarrow F$. {Thus $E \sim F$ can be verbalized as ' E is bijective with F '; the term ' E and F are *equipollent*' is also used.} Convention: $\emptyset \sim \emptyset$. The negation of $E \sim F$ is written $E \not\sim F$.

We are not prepared to say 'how many' elements a set E has, but if $E \sim F$ we can at least say that E and F have 'the same number' of elements. By our convention that functions are nonempty relations (1.3.1), $E \sim \emptyset$ and $\emptyset \sim E$ are true (by convention) only when $E = \emptyset$.

1.12.2. *Theorem.* For sets E, F, G ,

- (i) $E \sim E$;
- (ii) $E \sim F \Rightarrow F \sim E$;
- (iii) $E \sim F \ \& \ F \sim G \Rightarrow E \sim G$.

Proof. When $E = \emptyset$, all expressions $A \sim B$ in sight reduce to the convention $\emptyset \sim \emptyset$. Let us assume $E \neq \emptyset$.

- (i) The identity mapping $E \rightarrow E$ is a bijection.
- (ii) The inverse of a bijective mapping is bijective.
- (iii) The composite of two bijections is a bijection. \diamond

1.12.3. *Remark.* If \mathcal{S} is any set of sets, $E \sim F$ defines an equivalence relation in \mathcal{S} .¹

1.12.4. *Definition.* For sets E and F , if there exists an injection $E \rightarrow F$ we write $E \preceq F$ (or $F \succeq E$) and we say (just to have a way to verbalize the symbol) that E is **dominated** by F (or that F dominates E). Convention: $\emptyset \preceq E$ for every set E .

Another way to put the matter: $E \preceq F \Leftrightarrow E \sim A$ for some $A \subset F$. Note that $E \preceq \emptyset$ is true only when $E = \emptyset$ (and then, by convention).

1.12.5. *Theorem.* For sets E, F, G ,

- (i) $E \sim F \Rightarrow E \preceq F$; in particular, $E \preceq E$;
- (ii) $E \preceq F \ \& \ F \preceq G \Rightarrow E \preceq G$.

Proof. (i) If $E = \emptyset$ or $F = \emptyset$, the assertions reduce to notational conventions; otherwise, the implication reduces to the observation that a bijective mapping is injective.

(ii) Apart from conventions about the empty set, this reduces to the observation that the composite of injections is an injection. \diamond

The preceding theorem shows that in any set \mathcal{S} of sets, $E \preceq F$ defines a pre-order relation (1.7.1); the equivalence relation deduced from it (as in 1.7.7) is the relation of equipollence:

1.12.6. *Theorem.* (Schröder-Bernstein)² If $E \preceq F$ and $F \preceq E$ then $E \sim F$.

Proof. Given injective mappings $f : E \rightarrow F$ and $g : F \rightarrow E$, we have to come up with a bijection $E \rightarrow F$. It will suffice to prove that there exists a subset A of E such that

$$(*) \quad g(F - f(A)) = E - A;$$

the required bijection can then be pieced together from the restriction of f to A and the inverse of the restriction of g to $F - f(A)$ (Fig. 8).

In fact we will show that a set satisfying (*) exists for *every* pair of functions $f : E \rightarrow F$ and $g : F \rightarrow E$ (not necessarily injective). For

¹ The locution ‘the relation \sim is an equivalence relation in \mathcal{S} ’ was avoided because, technically, a ‘relation’ was defined to be a subset of a cartesian product of sets (1.2.2); there is no ‘universal set’ of sets in which \sim can be defined ‘once and for all’ (1.1.10). Now that we have noted what the locution does not say, there is no harm in using it.

² Cf. I. Kaplansky, op. cit., p. 33.

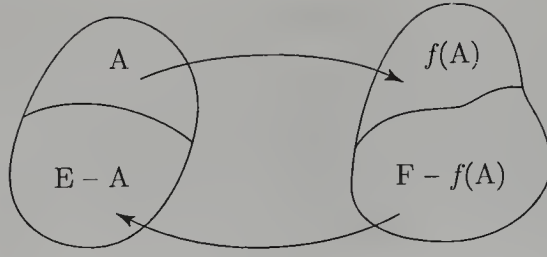


Figure 8

subsets $A \subset E$ and $B \subset F$, we write $\mathbb{C}A = E - A$ and $\mathbb{C}B = F - B$; the equation (*) then reads $g(\mathbb{C}f(A)) = \mathbb{C}A$, in other words,

$$(**) \quad \mathbb{C}g(\mathbb{C}f(A)) = A.$$

This looks like a 'fixed point' situation: for any $A \subset E$ we perform the operations indicated on the left side of (**); we are looking for an A that remains unchanged. At issue is the function $\varphi : \mathcal{P}(E) \rightarrow \mathcal{P}(E)$ defined by the formula

$$\varphi(S) = \mathbb{C}g(\mathbb{C}f(S)) \text{ for all } S \subset E.$$

With $\mathcal{P}(E)$ ordered by inclusion, φ is order-preserving:

$$S \subset T \Rightarrow \varphi(S) \subset \varphi(T).$$

{Reason: Mappings such as $S \mapsto f(S)$ preserve inclusion, whereas complementation reverses inclusion.} Let

$$\mathcal{S} = \{S \in \mathcal{P}(E) : \varphi(S) \supset S\}$$

(for example, $\emptyset \in \mathcal{S}$). Note that \mathcal{S} is invariant under φ : if $\varphi(S) \supset S$ then $\varphi(\varphi(S)) \supset \varphi(S)$, thus $S \in \mathcal{S} \Rightarrow \varphi(S) \in \mathcal{S}$. Let

$$A = \bigcup \mathcal{S} = \bigcup \{S : S \in \mathcal{S}\};$$

we will show that $\varphi(A) = A$. For all $S \in \mathcal{S}$ we have $A \supset S$, therefore $\varphi(A) \supset \varphi(S) \supset S$; thus $\varphi(A) \supset \bigcup \mathcal{S} = A$. This shows that $A \in \mathcal{S}$, therefore also $\varphi(A) \in \mathcal{S}$, whence $\varphi(A) \subset \bigcup \mathcal{S} = A$. \diamond

1.12.7. Corollary. If \mathcal{S} is any set of sets, $E \sim F$ is the equivalence relation in \mathcal{S} defined by equipollence, and $[E]$ denotes the equivalence class of a set $E \in \mathcal{S}$ (so that $[E] = [F] \Leftrightarrow E \sim F$), then the relation $[E] \leq [F]$ defined in the quotient set E/\sim by $E \preceq F$ is a partial ordering of E/\sim .

Proof. Immediate from 1.7.7, 1.12.5 and the theorem. \diamond

Suppose $E \sim F$. By 1.9.3, E is finite $\Leftrightarrow F$ is finite. More precisely, if $E \sim \mathbb{P}_n$ for some n then $F \sim \mathbb{P}_n$ and $\text{card } E = n = \text{card } F$ (1.9.14);

and if $E = \emptyset$ then $F = \emptyset$ and $\text{card } E = 0 = \text{card } F$. Let's extend this language to arbitrary sets:

1.12.8. *Definition.* If $E \sim F$ we say that E and F have the **same cardinality** and we write $\text{card } E = \text{card } F$.

What has been defined? A term ('same cardinality') and a notation ($\text{card } E = \text{card } F$). Implicit in this notation is that we are associating with each set E a symbol $\text{card } E$, and agreeing that two such symbols can be separated by the symbol "=" if and only if the sets to which the symbols correspond may be separated by the symbol " \sim " for equipollence.³ In 1.9.14, $\text{card } \mathbb{P}_n$ is in effect abbreviated to n .

1.12.9. *Definition.* Symbols $\text{card } E$ (and abbreviations for them) are called **cardinal numbers**. If u is a cardinal number (that is, u abbreviates $\text{card } F$ for some set F) and E is a set with $\text{card } E = u$ (in other words, $E \sim F$), we say that E **has cardinality** u (or that u is the cardinal number of E).

Thus, a set of cardinal numbers is the set of symbols $\text{card } E$ associated with some set S of sets E (however, if $E \in S$ and $F \sim E$, S is not obligated to contain F).

1.12.10. *Theorem.* If u, v, w are cardinal numbers, then

- (i) $u = u$;
- (ii) $u = v \Rightarrow v = u$;
- (iii) $u = v \ \& \ v = w \Rightarrow u = w$.

Proof. This is a rerun of 1.12.2. \diamond

Thus, in any set S of cardinal numbers, the relation $u = v$ introduced in 1.12.8 and 1.12.9 is an equivalence relation. Here are two especially useful notations (abbreviations, from our perspective):

1.12.11. *Definition.* We write \aleph_0 for $\text{card } \mathbb{P}$, and c for $\text{card } \mathbb{R}$ (called the *cardinal of the continuum*).

The symbol \aleph_0 is read "aleph nought" (aleph is the first letter of the Hebrew alphabet). The reason for the subscript zero is that \aleph_0 is, in an appropriate sense, the first infinite cardinal number (more about this in the present section—see (iv) of 1.12.21—and in the next two sections). The only thing new in the following theorem is the language of cardinality:

1.12.12. *Theorem.* A set E is denumerable if and only if its cardinality is \aleph_0 .

There is no proof for the preceding "theorem"; this is just Definition 1.10.5 in a new suit of clothes. So far we have just wrapped the language of

³This is called the "axiom for cardinal numbers" in the book of P. Suppes [*Axiomatic set theory*, Van Nostrand, Princeton, 1960; reprinted Dover, New York, 1972].

cardinality around the relation of equipollence (\sim); now let's do the same for the relation of domination (\preceq):

1.12.13. *Definition.* If $E \preceq F$ we write $\text{card } E \leq \text{card } F$ (or $\text{card } F \geq \text{card } E$).

Note that $u \geq 0$ for every cardinal number u ; for, if $u = \text{card } E$ then $\emptyset \preceq E$ by convention. In any set of cardinal numbers, the relation \leq is a partial ordering:

1.12.14. *Theorem.* If u, v, w are cardinal numbers, then

(i) $u = v \Rightarrow u \leq v$; in particular, $u \leq u$;

(ii) $u \leq v \ \& \ v \leq w \Rightarrow u \leq w$;

(iii) $u \leq v \ \& \ v \leq u \Rightarrow u = v$.

Proof. (i), (ii) A rerun of 1.12.5.

(iii) This is the Schröder-Bernstein theorem (1.12.6). \diamond

1.12.15. *Definition.* For cardinal numbers u and v we write $u < v$ (as in 1.7.5) if $u \leq v$ and $u \neq v$ (that is, $u \leq v$ is true and $u = v$ is false). The relation $u < v$ is also written $v > u$.

From this definition and (i) of 1.12.14, it is immediate that

$$u \leq v \Leftrightarrow u < v \text{ or } u = v$$

(that is, “ \leq ” is correctly verbalized as “less than or equal to”); the definition also rules out $u < u$. If $u = \text{card } E$ and $E \neq \emptyset$, then $u > 0$; for, $\emptyset \preceq E$ by convention but $\emptyset \sim E$ is impossible (1.12.1).

1.12.16. *Example.* $\aleph_0 < c$. For, $\aleph_0 \leq c$ (because $\mathbb{P} \subset \mathbb{R}$), but $\mathbb{P} \sim \mathbb{R}$ is ruled out by 1.10.11. We shall note in the next section that this is a special case of the following theorem:

1.12.17. *Theorem.* (Cantor) For every set E , $\text{card } E < \text{card } \mathcal{P}(E)$.

Proof. If $E = \emptyset$ then $\mathcal{P}(E) = \{\emptyset\}$ and the asserted inequality is obvious. If $E \neq \emptyset$ then $x \mapsto \{x\}$ is an injection $E \rightarrow \mathcal{P}(E)$, but $E \not\sim \mathcal{P}(E)$ by 1.10.6. \diamond

1.12.18. *Corollary.* If $(u_i)_{i \in I}$ is any family of cardinal numbers, then there exists a cardinal number u such that $u_i < u$ for all $i \in I$.

Proof. If $u_i = \text{card } E_i$ ($i \in I$), let $E = \bigcup E_i$; for all $i \in I$,

$$u_i = \text{card } E_i \leq \text{card } E < \text{card } \mathcal{P}(E),$$

thus $u = \text{card } \mathcal{P}(E)$ meets the requirements. \diamond

Eventually (§1.14) we will show that every set of cardinal numbers is well-ordered (in the sense of 1.11.2) by the relation \leq . In view of Zermelo's theorem (1.11.5), it is not surprising that the Axiom of Choice will figure in the proof; in fact, it is already needed for the following weaker result:

1.12.19. Theorem. *If u and v are cardinal numbers, then either $u \leq v$ or $v \leq u$.*

Proof. Apart from the trivial case that $u = 0$ or $v = 0$, this is a rerun of 1.11.12: if $u = \text{card } E$ and $v = \text{card } F$, then either $E \preceq F$ or $F \preceq E$. \diamond

1.12.20. Corollary. (*Law of Trichotomy*) *If u and v are cardinal numbers, then exactly one of the following is true: $u < v$, $u = v$, $u > v$.*

Proof. We know from the theorem that either $u \leq v$ or $v \leq u$. If $u \neq v$, we conclude that either $u < v$ or $v < u$ (1.12.15). Thus, at least one of the three conditions $u = v$, $u < v$, $u > v$ holds. To prove that *only one* holds, we must show that any two of the conditions are incompatible.

The conditions $u = v$, $u < v$ are incompatible by the definition of $u < v$, and similarly for the conditions $u = v$, $u > v$. Finally, the conditions $u < v$ and $v < u$ are incompatible; for, if both held, then $u \leq v$ and $u \geq v$, therefore $u = v$ by the Schröder-Bernstein theorem (1.12.6), contrary to the definition of $u < v$. \diamond

The organizing power of these notations is illustrated by the following theorem, essentially a summary of earlier observations:

1.12.21. Theorem. *For a set E ,*

- (i) E is denumerable $\Leftrightarrow \text{card } E = \aleph_0$;
- (ii) E is countable $\Leftrightarrow \text{card } E \leq \aleph_0$;
- (iii) E is finite $\Leftrightarrow \text{card } E < \aleph_0$;
- (iv) E is infinite $\Leftrightarrow \text{card } E \geq \aleph_0$;
- (v) E is uncountable $\Leftrightarrow \text{card } E > \aleph_0$.

Proof. (i) Already noted in 1.12.12.

(ii) Both sides are trivial if $E = \emptyset$. Assume $E \neq \emptyset$. If $\text{card } E \leq \aleph_0$ then $E \sim A \subset \mathbb{P}$ for some A ; since \mathbb{P} is countable, so is its subset A (1.10.2), therefore so is E . Conversely, if E is countable then, by 1.10.4, either $E \sim \mathbb{P}$ or $E \sim \mathbb{P}_n \subset \mathbb{P}$ for some n , and in either case $\text{card } E \leq \aleph_0$.

(iii) In view of (ii), the right side of (iii) says that E is countable but $E \not\sim \mathbb{P}$, therefore E is finite by 1.10.4. Conversely, if E is finite then $\text{card } E \leq \aleph_0$ by (ii), and $E \sim \mathbb{P}$ is ruled out by the infiniteness of \mathbb{P} , therefore $\text{card } E < \aleph_0$.

(iv) By 1.9.4, E is infinite if and only if $\mathbb{P} \preceq E$.

(v) If E is uncountable then it is infinite, so $\text{card } E \geq \aleph_0$ by (iv); but $E \sim \mathbb{P}$ is ruled out by the countability of \mathbb{P} , so $\text{card } E > \aleph_0$. Conversely, if $\text{card } E > \aleph_0$ then $\mathbb{P} \preceq E$ but $\mathbb{P} \not\sim E$, that is, E is infinite but not denumerable; if E were countable then (being not denumerable) it would be finite by 1.10.4, a contradiction. \diamond

The proof of the preceding theorem could be shortened slightly, at the cost of using the Schröder-Bernstein theorem and the Axiom of Choice. For

example, the equivalence (iv) is the contrapositive form of the equivalence (iii); for, the left sides are each other's negations by definition (1.9.1), whereas the right sides are also each other's negations: the condition $\text{card } E \geq \aleph_0$ means that either $\text{card } E > \aleph_0$ or $\text{card } E = \aleph_0$ (as remarked in 1.12.15), so its negation is $\text{card } E < \aleph_0$ by trichotomy (1.12.20). Similarly, (v) is the contrapositive form of (ii). The proof of (ii) itself can be shortened: for $E \neq \emptyset$, the left side of (ii) says that there exists a surjection $\mathbb{P} \rightarrow E$, and the right side says that there exists an injection $E \rightarrow \mathbb{P}$; the two sides are therefore equivalent by 1.5.4 (whose proof requires the Axiom of Choice). The theorem suggests the following terminology:

1.12.22. *Definition.* A cardinal number u is said to be **finite** if $u < \aleph_0$, **infinite** if $u \geq \aleph_0$, **countable** if $u \leq \aleph_0$, and **uncountable** if $u > \aleph_0$.

The set $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ of nonnegative integers is a set of abbreviations for the finite cardinals $\text{card } \emptyset$ and $\text{card } \mathbb{P}_n$ ($n = 1, 2, 3, \dots$); we may simply say that \mathbb{N} is the set of finite cardinals.

Exercises

1. For nonempty sets E and F , $E \lesssim F$ if and only if there exists a surjection $E \rightarrow F$.

2. For a pair of nonempty sets E and F , there exists a bijection $E \rightarrow F$ if and only if there exist an injection $E \rightarrow F$ and a surjection $E \rightarrow F$.

1.13. Cardinal Arithmetic, the Continuum Hypothesis

If u and v are cardinal numbers, we propose to define uv , $u+v$ and v^u (in that order).

Say $u = \text{card } E$, $v = \text{card } F$. If u and v are finite, say $u = m = \text{card } \mathbb{P}_m$, $v = n = \text{card } \mathbb{P}_n$, we expect $uv = mn = \text{card}(\mathbb{P}_m \times \mathbb{P}_n)$. In general, if $E \sim E'$ and $F \sim F'$, then $E \times F \sim E' \times F'$, so the following definition is unambiguous (i.e., single-valued):

1.13.1. *Definition.* If $u = \text{card } E$ and $v = \text{card } F$, we define

$$uv = \text{card}(E \times F).$$

1.13.2. *Remarks.* (i) $uv = 0 \Leftrightarrow u = 0$ or $v = 0$. {Reason: $E \times F = \emptyset \Leftrightarrow E = \emptyset$ or $F = \emptyset$.}

(ii) $uv = vu$. {If E and F are nonempty then $E \times F \sim F \times E$ via the mapping $(x, y) \mapsto (y, x)$.}

(iii) $(uv)w = u(vw)$. {For nonempty sets E, F, G , $(E \times F) \times G \sim E \times (F \times G)$ via $((x, y), z) \mapsto (x, (y, z))$.}

(iv) If $u \leq v$ and $u' \leq v'$, then $uu' \leq vv'$.

- (v) If $v > 0$ then $uv \geq u$. {If $b \in F$ then $E \lesssim E \times F$ via the mapping $x \mapsto (x, b)$.}
- (vi) $\aleph_0 \aleph_0 = \aleph_0$. { $\mathbb{P} \times \mathbb{P} \sim \mathbb{P}$ by 1.10.8. }

If E and F are finite and $E \cap F = \emptyset$, we expect $\text{card}(E \cup F) = \text{card } E + \text{card } F$. If E and F are arbitrary sets then, writing $E' = E \times \{0\}$ and $F' = F \times \{1\}$, we have $E \sim E'$, $F \sim F'$ and $E' \cap F' = \emptyset$.

1.13.3. *Definition.* Given cardinals u and v there exist, by the preceding remark, sets E' and F' with $u = \text{card } E'$, $v = \text{card } F'$ and $E' \cap F' = \emptyset$; define

$$u + v = \text{card}(E' \cup F').$$

If also $u = \text{card } E''$, $v = \text{card } F''$ and $E'' \cap F'' = \emptyset$, then $E' \cup F' \sim E'' \cup F''$, therefore $u + v$ is well-defined.

- 1.13.4. *Remarks.* (i) $u + v = 0 \Leftrightarrow u = 0 \ \& \ v = 0$.
(ii) $u + v = v + u$.
(iii) $(u + v) + w = u + (v + w)$.
(iv) $u \leq u + v$.
(v) If $u \leq v$ and $u' \leq v'$ then $u + u' \leq v + v'$.
(vi) $u(v + w) = uv + uw$. {If F and G are disjoint, then $E \times (F \cup G) = (E \times F) \cup (E \times G)$ with $E \times F$ and $E \times G$ disjoint.}
(vii) $\aleph_0 + \aleph_0 = \aleph_0$. {If A (resp. B) is the set of even (resp. odd) positive integers, then $\mathbb{P} = A \cup B$, $A \cap B = \emptyset$ and $\mathbb{P} \sim A \sim B$.}

The principle in cardinal arithmetic is that the big fish swallow the little ones; the following example is superseded by 1.13.7:

1.13.5. *Lemma.* If u is an infinite cardinal then $u + v = u$ for every countable cardinal v .

Proof. Say $u = \text{card } E$, $v = \text{card } F$, $E \cap F = \emptyset$; assuming E infinite and F countable, the problem is to show that $E \cup F \sim E$. Let D be a denumerable subset of E (1.9.4). Since F is either finite or denumerable (1.10.4) we can write $D = F' \cup D'$ with $F \sim F'$, $F' \cap D' = \emptyset$ and D' denumerable. Putting together the relations $F \sim F'$, $D \sim D'$, $E - D = E - D$, we have

$$F \cup D \cup (E - D) \sim F' \cup D' \cup (E - D),$$

that is, $F \cup E \sim E$. \diamond

1.13.6. *Theorem.* For every infinite cardinal u , $u + u = u$.

*Proof.*¹ Say $u = \text{card } E$. Since

$$E \times \{0, 1\} = (E \times \{0\}) \cup (E \times \{1\}),$$

we have $\text{card}(E \times \{0, 1\}) = u + u$.

We consider injective functions $f : A \rightarrow E \times \{0, 1\}$ defined on subsets A of E , such that $f(A) = A \times \{0, 1\}$ (which implies that $\text{card } A = \text{card } A + \text{card } A$). For example, if A is a denumerable subset of E , then

$$\text{card}(A \times \{0, 1\}) = \text{card } A + \text{card } A = \aleph_0 + \aleph_0 = \aleph_0 = \text{card } A$$

shows that $A \sim A \times \{0, 1\}$ and proves the existence of such a function f .

Let \mathcal{F} be the set of all such functions f and let \mathcal{G} be the set of all graphs G of such functions. As in 1.11.10, partially order \mathcal{F} by extension (this corresponds to the ordering of \mathcal{G} by inclusion). With upper bounds for chains constructed as in 1.11.11, \mathcal{F} is inductive; the crux of the matter is that if $f_i : A_i \rightarrow E \times \{0, 1\}$ is a family of pairwise comparable (one an extension of the other) functions in \mathcal{F} , $A = \bigcup A_i$, and $f : A \rightarrow E \times \{0, 1\}$ is defined so that $f|_{A_i} = f_i$ for all i , then

$$f(A) = \bigcup f(A_i) = \bigcup (A_i \times \{0, 1\}) = \left(\bigcup A_i \right) \times \{0, 1\} = A \times \{0, 1\},$$

thus $f \in \mathcal{F}$.

By Zorn's lemma, \mathcal{F} has a maximal element $f : A \rightarrow E \times \{0, 1\}$. We assert that $E - A$ is finite. If, on the contrary, $E - A$ is infinite, then there exist a denumerable subset D of $E - A$ and an injective mapping $g : D \rightarrow (E - A) \times \{0, 1\}$ such that $g(D) = D \times \{0, 1\}$. Then f and g can be combined to define a common extension $h : A \cup D \rightarrow E \times \{0, 1\}$ such that

$$h(A \cup D) = f(A) \cup g(D) = (A \times \{0, 1\}) \cup (D \times \{0, 1\}) = (A \cup D) \times \{0, 1\};$$

thus $h \in \mathcal{F}$ and, since h is a proper extension of f , the maximality of f is contradicted.

Since $E = A \cup (E - A)$ with $E - A$ finite, $E \sim A$ by the lemma; then

$$E \sim A \sim f(A) = A \times \{0, 1\} \sim E \times \{0, 1\},$$

so that $u = u + u$ by the first paragraph of the proof. \diamond

1.13.7. Corollary. If u and v are cardinals with u infinite and $v \leq u$, then $u + v = u$.

Proof. We have $u \leq u + v \leq u + u = u$ (by the theorem), therefore $u = u + v$ by the Schröder-Bernstein theorem ((iii) of 1.12.14). \diamond

Since any two cardinals are comparable (1.12.19), it follows that if one of u, v is infinite, then $u + v = \max\{u, v\}$.

¹ I learned this elegant proof from the book of Hewitt and Stromberg [*Real and abstract analysis*, Springer-Verlag, New York, 1965], p. 25, (4.2.9).

1.13.8. Theorem. *If u is an infinite cardinal, then $uu = u$.*

Proof. Say $u = \text{card } E$. We are interested in subsets A of E such that $A \sim A \times A$. Every denumerable subset of E has this property (1.10.8) and we would like to show that E itself has it.

Let \mathcal{F} be the set of all injective functions $f : A \rightarrow E \times E$, defined on subsets A of E , such that $f(A) = A \times A$ (and therefore $A \sim A \times A$); our problem is to show that there exists such a function with $A = E$. Partially order \mathcal{F} by extension; with upper bounds for chains constructed as in the proof of 1.13.6, \mathcal{F} is inductive. Here, the crux of the matter is that if $f_i : A_i \rightarrow E \times E$ is a pairwise comparable family in \mathcal{F} , $A = \bigcup A_i$ and $f : A \rightarrow E \times E$ is defined so that $f|_{A_i} = f_i$ for all i , then, since the sets A_i are comparable,

$$\begin{aligned} A \times A &= \bigcup_{i,j} (A_i \times A_j) = \bigcup_i (A_i \times A_i) \\ &= \bigcup_i f_i(A_i) = \bigcup_i f(A_i) = f(A), \end{aligned}$$

thus $f \in \mathcal{F}$.

By Zorn's lemma, \mathcal{F} has a maximal element $f : A \rightarrow E \times E$. Let $v = \text{card } A$. Since $vv = v$ (because $A \sim A \times A$), it will suffice to show that $u = v$. Note that A is infinite. For, if A were finite then $vv = v$ would imply that $v = 1$, $A = \{a\}$ for some point a ; the correspondence $f : a \mapsto (a, a)$ could be extended to a bijection $D \rightarrow D \times D$ with D denumerable (start with a denumerable set $D \subset E$ containing a , and any bijection $D \rightarrow D \times D$, then revise it—via a transposition—so that $a \mapsto (a, a)$), which would contradict the maximality of f .

We assert that $\text{card}(E - A) < v = \text{card } A$; this will complete the proof, since then

$$u = \text{card } E = \text{card } A + \text{card}(E - A) = v + \text{card}(E - A) = v$$

(the last equality by 1.13.7). Assume to the contrary (1.12.20) that $\text{card } A \leq \text{card}(E - A)$, so that $A \sim B \subset E - A$ for suitable B ; we will extend f to a bijection $A \cup B \sim (A \cup B) \times (A \cup B)$, thereby contradicting the maximality of f . Now, f already maps A to $A \times A$, so the task is to map B bijectively onto the set

$$S = (A \times B) \cup (B \times A) \cup (B \times B).$$

The possibility of doing this is a matter of sheer cardinality; let us compute. We have $v = \text{card } A = \text{card } B$ and $vv = v$, therefore

$$\text{card}(A \times B) = (\text{card } A)(\text{card } B) = vv = v$$

and similarly $\text{card}(B \times A) = \text{card}(B \times B) = v$. Thus S has cardinality $v + v + v = v$ (1.13.6), and so does B , therefore $B \sim S$. \diamond

1.13.9. Corollary. If u and v are nonzero cardinals, at least one of which is infinite, then $uv = \max\{u, v\}$.

Proof. Suppose, for example, that $0 < v \leq u$ (so that u is guaranteed to be infinite); then $u \leq uv \leq uu = u$ (by the theorem), therefore (Schröder-Bernstein) $uv = u = \max\{u, v\}$. \diamond

Cardinal arithmetic is certainly bizarre: $uv = u + v = \max\{u, v\}$ for any two nonzero cardinals at least one of which is infinite (1.13.9 and the remark following 1.13.7). The message is that if we want to make larger cardinals, we will have to work harder than just taking sums and products; exponentiation does the trick:

1.13.10. Definition. If E and F are nonempty sets, the set of all functions $f : E \rightarrow F$ is denoted $\mathcal{F}(E, F)$ or (for the reason that follows) F^E .

1.13.11. Remarks. (i) If E and F are finite, say $\text{card } E = m$ and $\text{card } F = n$, then $\mathcal{F}(E, F)$ has n^m elements (a function $f : E \rightarrow F$ can send each of the m points of E to any of the n points of F), whence the notation F^E .

(ii) If $f : E \rightarrow \{0, 1\}$ and $A = \{x \in E : f(x) = 1\}$, then f is the characteristic function φ_A of A ; thus $\{0, 1\}^E$ is the set of all characteristic functions of subsets of E , and $\mathcal{P}(E) \sim \{0, 1\}^E$ via the correspondence $A \mapsto \varphi_A$. For this reason, the power set $\mathcal{P}(E)$ is sometimes written 2^E .

(iii) For nonempty sets, if $E \sim E'$ and $F \sim F'$, then $F^E \sim F'^{E'}$. This justifies the following definition:

1.13.12. Definition. Let u and v be nonzero cardinal numbers, say $u = \text{card } E$ and $v = \text{card } F$; we define $v^u = \text{card}(F^E)$. Thus,

$$\text{card}(F^E) = (\text{card } F)^{\text{card } E}.$$

Conventions: $0^0 = v^0 = 1$ and $0^u = 0$ (recall that $u \neq 0$). {Rationale (cf. 1.5.5): The only relation $\emptyset \subset \emptyset \times F$ from \emptyset to F has domain \emptyset , but the only relation $\emptyset \subset E \times \emptyset$ from E to \emptyset does not have domain E ; the number of qualifying 'functions' is 1 and 0, respectively.}

1.13.13. Examples. (i) If $u = \text{card } E$, then $\text{card } \mathcal{P}(E) = 2^u$ (even if $u = 0$). {See 1.13.11, (ii).}

(ii) (Cantor's theorem) $u < 2^u$ for every cardinal u (1.12.17).

(iii) $v^{u+w} = v^u v^w$. {For nonempty sets E, F, G , with $E \cap G = \emptyset$, every function $f : E \cup G \rightarrow F$ corresponds to a pair (g, h) of functions $g : E \rightarrow F$, $h : G \rightarrow F$ via the formulas $g = f|E$, $h = f|G$.}

(iv) $(v^u)^w = v^{uw}$. {For nonempty sets E, F, G , every function $f : G \rightarrow \mathcal{F}(E, F)$ corresponds to a function $g : E \times G \rightarrow F$ via the formula $g(x, z) = [f(z)](x)$ (the value of the function $f(z) : E \rightarrow F$ at the point $x \in E$).}

1.13.14. Theorem. For nonempty sets E and F , F^E is the product of $\text{card } E$ copies of F , in the sense that there is a natural bijection

$$F^E \sim \prod_{x \in E} F.$$

Proof. On the right side, we are considering the family of sets $(F_x)_{x \in E}$ with $F_x = F$ for all $x \in E$. An element $f : E \rightarrow F$ of the left side corresponds to the point $(f(x))_{x \in E}$ of the product set on the right. \diamond

When $F = \{0, 1\}$, this yields the formula

$$2^E \sim \prod_{x \in E} \{0, 1\}.$$

1.13.15. Theorem. $2^{\aleph_0} = c$.

Proof. By the preceding formula (with $E = \mathbb{P}$), 2^{\aleph_0} is the cardinal number of the set of all sequences (x_1, x_2, x_3, \dots) with $x_n = 0$ or 1 for all $n \in \mathbb{P}$.

We show first that $[0, 1] \sim \mathbb{R}$, that is, the cardinality of the closed interval $[0, 1]$ is c . Since $(0, 1) \sim (-1, 1)$ (via the mapping $x \mapsto 2x - 1$) and $(-1, 1) \sim \mathbb{R}$ (via the mapping $x \mapsto x(1 - x^2)^{-1}$), we have $(0, 1) \sim \mathbb{R}$ and need only show that $[0, 1] \sim (0, 1)$; the latter follows from

$$[0, 1] \sim [\tfrac{1}{3}, \tfrac{1}{2}] \subset (0, 1) \subset [0, 1]$$

and the Schröder-Bernstein theorem.

We are now reduced to proving that $2^{\aleph_0} = \text{card } [0, 1]$. Define

$$\varphi : \{0, 1\}^{\mathbb{P}} \rightarrow [0, 1]$$

by the formula

$$\varphi(x_1, x_2, x_3, \dots) = \sum_{n=1}^{\infty} x_n / 2^n;$$

φ is surjective because every real number has a dyadic (base 2) representation, thus $2^{\aleph_0} \geq c$ by 1.5.4. Now define

$$\psi : \{0, 1\}^{\mathbb{P}} \rightarrow [0, 1]$$

by the formula

$$\psi(x_1, x_2, x_3, \dots) = \sum_{n=1}^{\infty} x_n / 3^n;$$

ψ is injective because the triadic (base 3) representation of a real number using only 0's and 1's (if possible at all) is unique, thus $2^{\aleph_0} \leq c$. By the Schröder-Bernstein theorem, $2^{\aleph_0} = c$. \diamond

In view of Cantor's theorem (1.13.13, (ii)), the formula $2^{\aleph_0} = c$ yields another proof of $\aleph_0 < c$ (1.12.16).

Does there exist a cardinal u such that $\aleph_0 < u < c$? The proposition that the answer is "no" is called the **Continuum Hypothesis**.²

(CH) If u is a cardinal number such that $\aleph_0 \leq u \leq 2^{\aleph_0}$ then either $u = \aleph_0$ or $u = 2^{\aleph_0}$.

By theorems of K. Gödel and P. Cohen, the Continuum Hypothesis is a bona fide axiom of set theory, consistent with and independent of ZF + AC (that is, neither disprovable nor provable using the Zermelo-Fraenkel axioms and the Axiom of Choice). The status of the **Generalized Continuum Hypothesis** is similar:

(GCH) If \aleph is an infinite cardinal number and u is a cardinal number such that $\aleph \leq u \leq 2^\aleph$, then either $u = \aleph$ or $u = 2^\aleph$.

No applications of (GC) or (GCH) are made in this text, though their relevance to the material in the next section will be evident.³

Exercises

1. Let u be a cardinal number and let $X = \bigcup_{i \in I} X_i$, where $(X_i)_{i \in I}$ is a family of sets such that $\text{card } X_i \leq u$ for all $i \in I$. Then:

(i) $\text{card } X \leq (\text{card } I)u$.

(ii) If u is infinite and $\text{card } I \leq u$, then $\text{card } X \leq u$.

(iii) In particular, a countable union of countable sets is countable.

{Hint: Let T be a set with $\text{card } T = u$ and, for each $i \in I$, let $f_i : X_i \rightarrow T$ be an injection. Let $Y = \bigcup_{i \in I} (\{i\} \times X_i) \subset I \times X$. There exist a surjection $Y \rightarrow X$ and an injection $Y \rightarrow I \times T$.}

2. If w is an infinite cardinal, then $c^w = 2^w$. In particular, $c^{\aleph_0} = c$ and $c^c = 2^c$.

{Hint: 1.13.15, 1.13.13, (iv) and 1.13.9.}

3. Let X be a set and let $S = \mathcal{F}(\mathbb{P}, X)$ be the set of all sequences (x_n) with $x_n \in X$ for all $n \in \mathbb{P}$. If $\text{card } X \leq 2^u$, where u is an infinite cardinal, then also $\text{card } S \leq 2^u$. In particular, if $\text{card } X \leq c$ then also $\text{card } S \leq c$.

{Hint: Exercise 2.}

² For a discussion of the history of such matters, see the book of I. Kaplansky [*Set theory and metric spaces*, 2nd edn., Chelsea, New York, 1977], §3.4.

³ For some interesting applications, see the books of L. Gillman and M. Jerison [*Rings of continuous functions*, Van Nostrand, Princeton, 1960; reprinted Springer, New York, 1976] and J. Oxtoby [*Category and measure*, Springer, New York, 1971].

1.14. Ordinality

We review several key definitions from §1.7. If E and F are pre-ordered sets, a mapping $f : E \rightarrow F$ is said to be an **order morphism** if $x \leq x' \Rightarrow f(x) \leq f(x')$, an **order monomorphism** if it is injective and satisfies $x \leq x' \Leftrightarrow f(x) \leq f(x')$, and an **order isomorphism** if it is a bijective order monomorphism. When E and F are simply ordered, ‘order monomorphism’ means the same as ‘injective order morphism’, and ‘order isomorphism’ the same as ‘bijective order morphism’ (1.7.17). If there exists an order isomorphism $E \rightarrow F$ we say that E and F are **similar** and we write $E \approx F$ (1.7.14).

In this section, our focus is on *well-ordered* sets. A key feature of the theory is that order morphisms are very much better behaved for well-ordered sets than for arbitrary (even simply) ordered sets; here is a striking illustration:

1.14.1. Theorem. *If E is a well-ordered set and $f : E \rightarrow E$ is an order monomorphism, then $f(x) \geq x$ for all $x \in E$.*

Proof. Assume to the contrary that the set $S = \{x \in E : f(x) < x\}$ is nonempty and let a be its smallest element. Since f is injective and preserves order, it follows from $f(a) < a$ that $f(f(a)) < f(a)$; thus $f(a)$ qualifies for membership in S , and $f(a) < a$ contradicts the minimality of a . \diamond

1.14.2. Corollary. *If E is a well-ordered set, then the only order isomorphism $E \rightarrow E$ is the identity mapping.*

Proof. If $f : E \rightarrow E$ is an order isomorphism then so is f^{-1} ; for every $x \in E$ we have $f(x) \geq x$ and $f^{-1}(x) \geq x$ by the theorem, therefore $f(x) = x$. \diamond

1.14.3. Corollary. *If E and F are well-ordered sets such that $E \approx F$, then there exists exactly one order isomorphism $f : E \rightarrow F$.*

Proof. To say that $E \approx F$ means that there is at least one order isomorphism $f : E \rightarrow F$; if $g : E \rightarrow F$ is another, then $g^{-1} \circ f = \text{id}_E$ by the preceding corollary. \diamond

Notationally, the discussion of ordinality parallels that of cardinality (§1.12):

1.14.4. Definition. Let E and F be well-ordered sets. If $E \approx F$ we say that E and F have the **same ordinality** and we write $\text{ord } E = \text{ord } F$.

Implicit in the notation is that we are associating with each well-ordered set E a symbol $\text{ord } E$, and agreeing to write $\text{ord } E = \text{ord } F$ precisely when E and F are *similar* well-ordered sets.

1.14.5. *Definition.* Symbols $\text{ord } E$ (and abbreviations for them) are called **ordinal numbers**. If α is an ordinal number (that is, α abbreviates $\text{ord } F$ for some well-ordered set F) and if E is a well-ordered set such that $\text{ord } E = \alpha$ (in other words, $E \approx F$), we say that E has **ordinality** α (or that α is the ordinal number of E). Convention: $\text{ord } \emptyset = 0$.

A set Λ of ordinal numbers is the set of symbols $\text{ord } E$ associated with some set \mathcal{S} of well-ordered sets E , that is, $\Lambda = \{\text{ord } E : E \in \mathcal{S}\}$.

1.14.6. *Theorem.* If α, β, γ are ordinal numbers, then

- (i) $\alpha = \alpha$;
- (ii) $\alpha = \beta \Rightarrow \beta = \alpha$;
- (iii) $\alpha = \beta \ \& \ \beta = \gamma \Rightarrow \alpha = \gamma$.

Proof. This is a restatement of the properties of \approx listed in 1.7.15, (i). \diamond

Thus, in any set of ordinal numbers, the relation $\alpha = \beta$ of 1.14.5 is an equivalence relation.

1.14.7. *Definition.* The set $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ of nonnegative integers is well-ordered for the usual ordering; its ordinal number is denoted ω ("little omega"), thus $\omega = \text{ord } \mathbb{N}$. {"Big omega" Ω will make its appearance later on (1.14.36).}

The comparison of ordinal numbers is more subtle than that of cardinal numbers. The basis for the definition of $\alpha < \beta$ is the following concept:

1.14.8. *Definition.* Let E be a well-ordered set, $a \in E$. The set

$$E(a) = \{x \in E : x < a\}$$

is called the **initial segment** of E determined by a .

1.14.9. *Remarks.* (i) $E(a) = \emptyset \Leftrightarrow a$ is the first element of E .

(ii) Let E be a well-ordered set. Every subset of E is well-ordered for the order relation it inherits from E . In particular, let $a \in E$ and regard $E' = E(a)$ as a well-ordered set. If $b \in E'$ (that is, $b < a$), then $E'(b) = E(b)$; expressed somewhat clumsily, $(E(a))(b) = E(b)$.

Continuing the list of consequences of Theorem 1.14.1:

1.14.10. *Corollary.* If E is a well-ordered set and $a \in E$, then there does not exist an order monomorphism $E \rightarrow E(a)$. In particular, E is not similar to any of its initial segments.

Proof. Assume to the contrary that, for some $a \in E$, there exists an order monomorphism $f : E \rightarrow E(a)$. Composing f with the insertion mapping $i : E(a) \rightarrow E$, we have an order monomorphism $i \circ f : E \rightarrow E$.

By 1.14.1, $(i \circ f)(x) \geq x$ for all $x \in E$; in particular, $f(a) \geq a$, contrary to $f(a) \in E(a)$. \diamond

1.14.11. Corollary. *If E is a well-ordered set and $x, y \in E$, then the following conditions are equivalent:*

- (a) $x \leq y$;
- (b) $E(x) \subset E(y)$;
- (c) *there exists an order monomorphism $E(x) \rightarrow E(y)$.*

Proof. The implications (a) \Rightarrow (b) \Rightarrow (c) are obvious.

(c) \Rightarrow (a): Let $f : E(x) \rightarrow E(y)$ be an order monomorphism and assume to the contrary that $y < x$; then $E(y) = (E(x))(y)$ and f is an order monomorphism of $E(x)$ into one of its initial segments, contrary to 1.14.10. \diamond

1.14.12. Corollary. *If E is a well-ordered set and $x, y \in E$, then the following conditions are equivalent: (a) $x = y$; (b) $E(x) = E(y)$; (c) $E(x) \approx E(y)$.*

Proof. The implications (a) \Rightarrow (b) \Rightarrow (c) are obvious. Assuming (c), we have $x \leq y$ and $y \leq x$ by “(c) \Rightarrow (a)” of the preceding corollary, thus $x = y$. \diamond

The preceding two corollaries focus attention on the set of all initial segments of a well-ordered set:

1.14.13. *Definition.* If E is a well-ordered set, we write $\mathcal{I}(E)$ for the set of all initial segments of E ,

$$\mathcal{I}(E) = \{E(x) : x \in E\}.$$

Convention: $\mathcal{I}(\emptyset) = \emptyset$.

With inclusion as the ordering, $\mathcal{I}(E)$ is a partially ordered set; in fact, it is well-ordered:

1.14.14. Corollary. *For every nonempty well-ordered set E , $E \approx \mathcal{I}(E)$ via the mapping $x \mapsto E(x)$.*

Proof. The mapping in question is surjective by the definition of $\mathcal{I}(E)$, injective by 1.14.12, and an order morphism by 1.14.11. \diamond

Given a pair of ordinal numbers $\alpha = \text{ord } E$ and $\beta = \text{ord } F$, it would be natural to write “ $\alpha \leq \beta$ ” if there exists an order monomorphism $E \rightarrow F$, and “ $\alpha < \beta$ ” in case “ $\alpha \leq \beta$ ” but $\alpha \neq \beta$ (that is, $E \not\approx F$). We shall see below (in 1.14.20) that this is equivalent to the following alternative definition:

1.14.15. *Definition.* Let α and β be ordinal numbers, say $\alpha = \text{ord } E$ and $\beta = \text{ord } F$. We write $\alpha < \beta$ (or $\beta > \alpha$) if $E \approx F(b)$ for some $b \in F$, and we write $\alpha \leq \beta$ (or $\beta \geq \alpha$) if either $\alpha < \beta$ or $\alpha = \beta$. Convention: $0 < \alpha$ for every nonzero ordinal number α (that is, when $E \neq \emptyset$); then $0 \leq \alpha$ for every ordinal number α .

If also $\alpha = \text{ord } E'$ and $\beta = \text{ord } F'$, so that $E \approx E'$ and $F \approx F'$, it is clear that E is similar to an initial segment of F if and only if E' is similar to an initial segment of F' , thus the relations $\alpha < \beta$ and $\alpha \leq \beta$ are well-defined. Also, the relations $\alpha < \beta$ and $\alpha = \beta$ are mutually exclusive, since $\alpha < \alpha$ is ruled out by 1.14.10.

We remark that if $\text{ord } E < \text{ord } F$ then $E \approx F(b)$ for a *unique* $b \in F$ (1.14.12), and there is only one order isomorphism implementing the similarity $E \approx F(b)$ (1.14.3).

1.14.16. *Lemma.* The relation $\alpha \leq \beta$ of 1.14.15 is a partial ordering of every set of ordinal numbers; that is, for ordinal numbers α, β, γ ,

- (i) $\alpha \leq \alpha$,
- (ii) $\alpha \leq \beta \ \& \ \beta \leq \gamma \Rightarrow \alpha \leq \gamma$,
- (iii) $\alpha \leq \beta \ \& \ \beta \leq \alpha \Rightarrow \alpha = \beta$.

Proof. (i) In fact $\alpha = \alpha$.

(ii) If $\alpha = 0$ or $\alpha = \beta$ or $\beta = \gamma$, the implication is obvious. Suppose $0 < \alpha$, $\alpha < \beta$ and $\beta < \gamma$; say $\alpha = \text{ord } E$, $\beta = \text{ord } F$, $\gamma = \text{ord } G$. By assumption, $E \approx F(b)$ and $F \approx G(c)$ for suitable elements b and c ; if d is the image of b under the order isomorphism that implements $F \approx G(c)$, then $d < c$ and $F(b) \approx (G(c))(d) = G(d)$, whence $E \approx G(d)$, so that $\alpha < \gamma$.

(iii) Let $\alpha \leq \beta$ and $\beta \leq \alpha$ and assume to the contrary that $\alpha \neq \beta$. Then $\alpha < \beta$ and $\beta < \alpha$, and the proof of (ii) yields the contradiction $\alpha < \alpha$. \diamond

In what follows, the only ordering contemplated on sets of ordinal numbers is the relation \leq of 1.14.15; we call it the **natural ordering** of ordinal numbers.

Our next goal is to show that every set of ordinal numbers is well-ordered for the natural ordering (in other words, in every set of well-ordered sets, there is one among them that is either similar to, or similar to an initial segment of, each of the others). It is useful to have a characterization of those subsets of a well-ordered set that are initial segments:

1.14.17. *Lemma.* Let E be a well-ordered set, $A \subset E$. The following conditions are equivalent:

- (a) A is an initial segment of E ;
- (b) $A \neq E$ and the relations $x \in A$, $y \in E$, $y \leq x$ imply that $y \in A$.

Proof. Note that (a) entails $E \neq \emptyset$, by the convention $\mathcal{I}(\emptyset) = \emptyset$.

(a) \Rightarrow (b): Obvious.

(b) \Rightarrow (a): Let z be the first element of the (nonempty) set $E - A$; it will suffice to show that $E - A = E - E(z)$. Since E is simply ordered, $E - E(z) = \{x \in E : x \geq z\}$. If $x \in E - A$ then $x \geq z$ by the choice of z , thus $E - A \subset E - E(z)$. Conversely, if $x \in E - E(z)$, that is, $x \geq z$, then $x \in E - A$; for, the alternative $x \in A$, coupled with $z \leq x$, would imply that $z \in A$ by (b), a contradiction. \diamond

It is convenient to have a name for the second property in (b) of the lemma:

1.14.18. *Definition.* A subset A of a pre-ordered set X is said to be an **ideal** (or 'order ideal') of X if it has the following property:

$$(x \in A, y \in X, y \leq x) \Rightarrow y \in A.$$

The preceding lemma says that the only ideals of a well-ordered set E are E itself and its initial segments. If $\alpha = \text{ord } E$ and $\beta = \text{ord } F$, then $\alpha \leq \beta$ means that E is similar to an ideal of F (1.14.15).

A major step on the way to proving that \leq is a well-ordering of every set of ordinals is to prove that it is a simple ordering:

1.14.19. *Theorem. (Trichotomy)* If α and β are ordinal numbers, then exactly one of the following three conditions holds: (1) $\alpha < \beta$, (2) $\alpha = \beta$, (3) $\beta < \alpha$.

Proof. Say $\alpha = \text{ord } E$, $\beta = \text{ord } F$, $\gamma = \text{ord } G$. The problem is to show that exactly one of the following conditions holds: (1) $E \approx F(b)$ for some $b \in F$, (2) $E \approx F$, or (3) $F \approx E(a)$ for some $a \in E$. If $E = F = \emptyset$ then only (2) holds; if $E = \emptyset$ and $F \neq \emptyset$ then only (1) holds; if $E \neq \emptyset$ and $F = \emptyset$ then only (3) holds. We can thus assume that E and F are both nonempty.

Since $\alpha < \alpha$ is ruled out, no two of the conditions (1), (2), (3) can hold simultaneously, thus *at most* one of them can hold; we must show that *at least* one of them holds. The strategy is to consider pairs of subsets $A \subset E$, $B \subset F$ such that $A \approx B$ and to show that one can arrange to have either $A = E$ or $B = F$ (or both). Let

$$A = \{x \in E : E(x) \approx F(y) \text{ for some } y \in F\},$$

$$B = \{y \in F : E(x) \approx F(y) \text{ for some } x \in E\};$$

since $\emptyset \approx \emptyset$ we are assured that A contains the first element of E , and B contains the first element of F . Note that if $E(x) \approx F(y)$ then $x \in A$, $y \in B$ and each of x and y uniquely determines the other (1.14.12); this means that the set $\{(x, y) : E(x) \approx F(y)\}$ is the graph of a bijection $f : A \rightarrow B$, characterized by the condition

$$(*) \quad E(x) \approx F(f(x)) \text{ for all } x \in A;$$

moreover, for each $x \in A$, the similarity in (*) is implemented by a unique order isomorphism (1.14.3).

We are going to show that either $A = E$ or $B = F$ (or both). First, we note that A and B are ideals. For example, assuming $x' < x \in A$, let us show that $x' \in A$. We have $x' \in E(x)$; if $y' \in F(f(x))$ is the image of x' under the order isomorphism that implements the similarity $E(x) \approx F(f(x))$, then $E(x') \approx F(y')$ via a restriction of this isomorphism; this shows that $x' \in A$ (and that $y' \in B$, $y' = f(x')$).

The preceding argument also shows that, for each $x \in A$, $E(x) \subset A$ (similarly $F(y) \subset B$ for each $y \in B$), the unique order isomorphism $E(x) \approx F(f(x))$ is the restriction of f to $E(x)$, and $f(E(x)) = F(f(x))$. It follows that f is an order isomorphism of A onto B . For, if $x, x' \in A$ and $x' < x$, then $f(x') \in f(E(x)) = F(f(x))$, therefore $f(x') < f(x)$; this shows that f is an injective order morphism, and f^{-1} takes care of itself (1.7.17). Thus f implements a similarity $A \approx B$.

Finally, we assert that $A = E$ or $B = F$. The alternative is that A and B are initial segments of E and F , respectively (1.14.17), say $A = E(x)$ and $B = F(y)$; then $E(x) = A \approx B \approx F(y)$, therefore $x \in A = E(x)$ yields the absurdity $x < x$. If $A = E$ then either (1) or (2) holds, and if $B = F$ then either (2) or (3) holds. \diamond

1.14.20. Corollary. *For nonempty well-ordered sets E and F ,*

$$\text{ord } E \leq \text{ord } F \Leftrightarrow \exists \text{ order monomorphism } E \rightarrow F.$$

Proof. \Rightarrow : Immediate from the definitions.

\Leftarrow : Assuming $g : E \rightarrow F$ is an order monomorphism (so that $E \approx g(E)$), the problem is to show that E is similar to an ideal of F (remarks following 1.14.18). Let A and B be the ideals and $f : A \rightarrow B$ the order isomorphism constructed in the proof of the preceding theorem; it will suffice to show that $A = E$. Assume to the contrary that $A = E(a)$ for some $a \in E$; as shown in the proof of the theorem, this implies that $B = F$. We then have the diagram

$$E \xrightarrow{g} F = B \xrightarrow{f^{-1}} A = E(a),$$

thus $f^{-1} \circ g : E \rightarrow E(a)$ is an order monomorphism, in contradiction to 1.14.10. \diamond

1.14.21. Corollary. *If E is a well-ordered set, then every subset of E is similar to an ideal of E .*

Proof. Let $A \subset E$. If $A = \emptyset$ the assertion is covered by the convention $\emptyset \approx \emptyset$. If $A \neq \emptyset$, apply the preceding corollary to the insertion mapping $A \rightarrow E$ to conclude that $\text{ord } A \leq \text{ord } E$. \diamond

The analogue of the Schröder-Bernstein theorem holds for order morphisms of well-ordered sets:

1.14.22. Corollary. *If E and F are well-ordered sets and if there exist order monomorphisms $E \rightarrow F$ and $F \rightarrow E$, then $E \approx F$.*

Proof. By 1.14.20, $\text{ord } E \leq \text{ord } F$ and $\text{ord } F \leq \text{ord } E$, therefore $\text{ord } E = \text{ord } F$ (1.14.16). \diamond

Note that the assertion of the corollary fails if “well-ordered” is replaced by “simply ordered” (1.7.15, (iii)).

We now elaborate on the similarity $E \approx \mathcal{I}(E)$ of 1.14.14 to show that certain very special sets of ordinal numbers are well-ordered for the natural ordering. We shall see later that there is no ‘set of all ordinal numbers’; however, for a fixed ordinal number α , the set of all ordinals β such that $\beta < \alpha$ makes perfect sense:

1.14.23. Lemma. *For every ordinal number α , there exists a unique set $I(\alpha)$ of ordinal numbers such that, for an ordinal number β ,*

$$(*) \quad \beta \in I(\alpha) \Leftrightarrow \beta < \alpha.$$

Moreover, $I(\alpha)$ is well-ordered and $\alpha = \text{ord } I(\alpha)$.

Proof. Let E be a well-ordered set such that $\text{ord } E = \alpha$.

Existence. Let

$$I(\alpha) = \{\text{ord } A : A \in \mathcal{I}(E)\} = \{\text{ord } E(x) : x \in E\}.$$

If $\beta \in I(\alpha)$, say $\beta = \text{ord } E(x)$, then $\beta < \alpha$ because $E(x)$ is (trivially) similar to the initial segment $E(x)$ of E . Conversely, if β is an ordinal with $\beta < \alpha$, say $\beta = \text{ord } F$ and $F \approx E(x)$, then $\beta = \text{ord } F = \text{ord } E(x) \in I(\alpha)$. Thus $I(\alpha)$ satisfies (*).

Uniqueness. If I is any set satisfying the analogue of (*), then

$$\beta \in I \Leftrightarrow \beta < \alpha \Leftrightarrow \beta \in I(\alpha),$$

thus $I = I(\alpha)$.

Finally, define $f : E \rightarrow I(\alpha)$ by $f(x) = \text{ord } E(x)$; f is surjective by the definition of $I(\alpha)$. If $f(x) = f(y)$, that is, $E(x) \approx E(y)$, then $x = y$ by 1.14.12, thus f is injective. Moreover, if $x, y \in E$ and $x < y$, then $E(x) = (E(y))(x)$ shows that $\text{ord } E(x) < \text{ord } E(y)$, thus f is an order morphism and therefore an order isomorphism (1.7.17). In particular, $I(\alpha)$ is well-ordered (for the natural ordering) and $E \approx I(\alpha)$, whence $\alpha = \text{ord } E = \text{ord } I(\alpha)$. \diamond

1.14.24. Theorem. *Every set of ordinal numbers is well-ordered for the natural ordering (1.14.15).*

Proof. It suffices to show that if Λ is any nonempty set of ordinal numbers, then Λ has a smallest element. Let $\alpha \in \Lambda$ and let $\Lambda' = \Lambda \cap I(\alpha) = \{\beta \in \Lambda : \beta < \alpha\}$. If $\Lambda' = \emptyset$ then α is the smallest element of Λ . Otherwise, Λ' is a nonempty subset of $I(\alpha)$; since $I(\alpha)$ is well-ordered by the lemma, Λ' has a smallest element γ . It follows that γ is also the

smallest element of Λ ; for, if $\beta \in \Lambda - \Lambda'$ then $\beta \geq \alpha$ (because Λ is simply ordered) and $\alpha > \gamma$ (because $\gamma \in \Lambda'$), therefore $\gamma < \beta$. \diamond

The preceding theorem, coupled with Zermelo's theorem, yields the corresponding theorem for cardinal numbers:

1.14.25. Corollary. *Every set of cardinal numbers is well-ordered for the relation \leq of 1.12.13.*

Proof. Given any nonempty set A of cardinal numbers, it suffices to show that A has a smallest element. Write A as an indexed family $A = \{u_i : i \in I\}$ (cf. 1.3.10). Say $u_i = \text{card } E_i$ ($i \in I$). If $E_i = \emptyset$ for some i , then $0 \in A$ and we are done. Assuming the E_i are all nonempty, let $E = \bigcup E_i$ be their union (1.4.1). By Zermelo's theorem (1.11.5) we can suppose that E is equipped with a well-ordering; then its subsets E_i are also well-ordered and we can consider the nonempty set

$$\Lambda = \{\text{ord } E_i : i \in I\}$$

of nonzero ordinal numbers. By the preceding theorem, Λ is well-ordered, thus there exists an index j such that $\text{ord } E_j \leq \text{ord } E_i$ for all i . This means that, for every i , there exists an order monomorphism $E_j \rightarrow E_i$, so in particular $\text{card } E_j \leq \text{card } E_i$; thus $u_j = \text{card } E_j$ is the smallest element of A . \diamond

It is easy to see that there is no 'largest ordinal': if α is any ordinal, say $\alpha = \text{ord } E$, adjoin to E an element z not in E , declare it to be greater than every element of E , and let $F = E \cup \{z\}$ be the resulting well-ordered set; then $E = F(z)$, therefore $\alpha < \text{ord } F$. {Alternatively, any well-ordering of $\mathcal{P}(E)$ is guaranteed to have ordinality $> \text{ord } E$ by a combination of Cantor's theorem (1.12.17) and the comparability of ordinals (1.14.19).} Here is a subtler result (that shuts the door on the existence of 'the set of all ordinal numbers'):

1.14.26. Theorem. *If Λ is any set of ordinal numbers, then there exists a smallest ordinal number β such that $\alpha < \beta$ for all $\alpha \in \Lambda$.*

Proof. Let

$$\Lambda^* = \Lambda \cup \bigcup_{\alpha \in \Lambda} I(\alpha).$$

It is clear that an ordinal number γ belongs to Λ^* if and only if $\gamma \leq \alpha$ for some $\alpha \in \Lambda$; it follows that, for every $\gamma \in \Lambda^*$, $I(\gamma) \subset \Lambda^*$ and therefore $I(\gamma) = \Lambda^*(\gamma)$. Equipped with the natural ordering, Λ^* is a well-ordered set (1.14.24); let $\beta = \text{ord } \Lambda^*$. For all $\alpha \in \Lambda$, we have $I(\alpha) = \Lambda^*(\alpha)$, therefore (cf. 1.14.23)

$$\alpha = \text{ord } I(\alpha) < \text{ord } \Lambda^* = \beta.$$

Finally, if δ is an ordinal number such that $\alpha < \delta$ for all $\alpha \in \Lambda$, we assert that $\beta \leq \delta$. Assume to the contrary that $\delta < \beta$. Say $\delta = \text{ord } D$, where $D \approx \Lambda^*(\gamma) = I(\gamma)$ for some $\gamma \in \Lambda^*$. Then $\gamma \leq \alpha$ for some $\alpha \in \Lambda$, therefore

$$\delta = \text{ord } D = \text{ord } I(\gamma) = \gamma \leq \alpha,$$

a contradiction. \diamond

In particular, given any ordinal number α , there is a smallest ordinal α' with $\alpha' > \alpha$ (α' is called the **successor** of α); it is plausible that α' is the ordinal $\text{ord}(E \cup \{z\})$ described following 1.14.25, and reasonable that it should be denoted $\alpha + 1$. Before entering into details, we must look after some notational matters.

1.14.27. *Definition.* An ordinal number α is said to be **finite** (**infinite**, **denumerable**, **countable**, **uncountable**) if $\alpha = \text{ord } E$ with E a finite (infinite, etc.) well-ordered set.

Since $E \approx F \Rightarrow E \sim F$, the foregoing concepts depend only on α , not the particular well-ordered set E of which it is the ordinal number. For example, $\omega = \text{ord } \mathbb{N}$ is infinite, denumerable, and countable. The initial segments of \mathbb{N} are its subsets $\mathbb{N}(n) = \{k \in \mathbb{N} : k < n\} = \{0, 1, \dots, n-1\}$; their ordinal numbers are the finite ordinals:

1.14.28. *Theorem.* For an ordinal number α ,

$$\alpha \text{ is finite} \Leftrightarrow \alpha < \omega.$$

Proof. We can suppose $\alpha > 0$. Say $\alpha = \text{ord } E$.

\Leftarrow : Since $0 < \alpha < \omega$ we have $E \approx \mathbb{N}(n)$ for some positive integer n ; in particular, $E \sim \mathbb{N}(n) \sim \mathbb{P}_n = \{1, \dots, n\}$, so E is finite with $\text{card } E = n$ (1.9.14).

\Rightarrow : By assumption, E is finite and nonempty, thus $E \sim \mathbb{N}(n)$ for some positive integer n . Define a function $f : \mathbb{N}(n) \rightarrow E$ recursively, as follows: $f(0)$ is the smallest element of E , and $f(k+1)$ is the smallest element of $E - \{f(0), \dots, f(k)\}$. Since $f(0) < f(1) < \dots < f(n-1)$, f is an order isomorphism, thus $E \approx \mathbb{N}(n)$ and $\alpha = \text{ord } E = \text{ord } \mathbb{N}(n) < \text{ord } \mathbb{N} = \omega$. \diamond

The foregoing proof shows that there can be no harm in abbreviating $\text{ord } \mathbb{N}(n)$ to n ; the theorem says that the set of all finite ordinals is $I(\omega)$ and its proof shows that (after the indicated abbreviation)

$$I(\omega) = \{0, 1, 2, \dots\}.$$

(Notation comes into the world destined to be recycled: we are using the same letter n to represent an integer, a finite cardinal and a finite ordinal.)

Before outright identifying $I(\omega)$ with \mathbb{N} , we must contemplate the addition of ordinal numbers and check that, for finite ordinals, there are no surprises:

1.14.29. *Definition.* If α and β are ordinal numbers, $\alpha + \beta$ is defined as follows: let E and F be well-ordered sets such that $\alpha = \text{ord } E$, $\beta = \text{ord } F$ and $E \cap F = \emptyset$ (cf. 1.13.2), extend the orderings of E and F to $E \cup F$ by declaring $x < y$ for all $x \in E$ and $y \in F$ (this yields a well-ordering of $E \cup F$) and define $\alpha + \beta = \text{ord}(E \cup F)$. Convention: $\alpha + 0 = 0 + \alpha = \alpha$.

Note that $\alpha + \beta$ depends only on α and β , not on the particular pair of disjoint well-ordered sets E, F with ordinalities α, β .

1.14.30. *Examples.* (i) If F is a well-ordered set having a largest element z , and if $\alpha = \text{ord}(F - \{z\}) = \text{ord } F(z)$, then $\alpha + 1 = \text{ord } F > \alpha$.

(ii) For ordinals α and β , $\alpha < \beta \Leftrightarrow \alpha + 1 \leq \beta$, thus $\alpha + 1$ is the smallest ordinal $> \alpha$. {Proof: Suppose $\alpha < \beta$, say $\alpha = \text{ord } E$, $\beta = \text{ord } F$ and $E \approx F(y)$; since $F(y) \cup \{y\}$ has largest element y , it is clear from (i) that $\alpha + 1 = \text{ord}(F(y) \cup \{y\}) \leq \text{ord } F = \beta$. The reverse implication is immediate from $\alpha < \alpha + 1$.}

(iii) For every finite ordinal n , $n + \omega = \omega$. {Proof: We can suppose $n > 0$. Let F be a well-ordered set with $\text{ord } F = n$ and $F \cap \mathbb{N} = \emptyset$, order $F \cup \mathbb{N}$ as indicated in 1.14.29, let $f : F \rightarrow \mathbb{N}(n)$ be any order isomorphism, and define an order isomorphism $g : F \cup \mathbb{N} \rightarrow \mathbb{N}$ by $g(x) = f(x)$ for $x \in F$ and $g(x) = x + n$ for $x \in \mathbb{N}$.}

(iv) $1 + \omega = \omega$ but $\omega + 1 > \omega$.

(v) For an ordinal α ,

$$\alpha \text{ is infinite} \Leftrightarrow \alpha \geq \omega,$$

thus ω is the smallest infinite ordinal; in particular, every infinite well-ordered set contains an order-theoretic copy of \mathbb{N} (also easy to see directly). {Proof: This is 1.14.28 in contrapositive form since, by the comparability of ordinals, the negation of $\alpha < \omega$ is $\alpha \geq \omega$.}

(vi) Recall that, for a cardinal number u ,

$$u \text{ uncountable} \Leftrightarrow u > \aleph_0,$$

where $\aleph_0 = \text{card } \mathbb{N}$ (1.12.21). In contrast, $\omega + 1 > \omega = \text{ord } \mathbb{N}$, but $\omega + 1$ is countable. Moral: The analogy between ordinals and cardinals should be approached with caution.

The preceding examples show that there are many denumerable ordinals, but ω is \leq each of them. To put it another way, there are many ways of well-ordering the set \mathbb{N} , but the usual ordering is minimal in the appropriate sense: \mathbb{N} equipped with the usual ordering is order-isomorphic to a subset of \mathbb{N} when \mathbb{N} is equipped with any other well-ordering. We shall see in the next theorem that this is a special case of a general principle; first, let us extend an idea broached in 1.14.27, by assigning to every ordinal number a cardinal number in the natural way:

1.14.31. *Definition.* If α is an ordinal number, say $\alpha = \text{ord } E$, we define the **cardinal number** of α , denoted $\text{card } \alpha$, to be $\text{card } E$.

If $E \approx F$ then $E \sim F$, so $\text{card } \alpha$ depends only on α , not on the particular well-ordered set of which α is the ordinal number. For example, $\text{card}(\omega + \omega) = \aleph_0$ and $\text{card } n = n$ for every finite ordinal n . There is likewise a natural way to assign an ordinal number to each cardinal:

1.14.32. Theorem. *For each cardinal number u , there exists a smallest ordinal number α such that $\text{card } \alpha = u$.*

Proof. We can suppose $u \neq 0$. Let E be a set with $\text{card } E = u$. Consider the set \mathcal{R} of all subsets $R \subset E \times E$ such that the relation xRy is a well-ordering of E ; by Zermelo's theorem, $\mathcal{R} \neq \emptyset$. For each $R \in \mathcal{R}$, let us write E_R for E equipped with the well-ordering defined by R . In effect, the set $\{E_R : R \in \mathcal{R}\}$ is a complete set of models for well-ordered sets F of cardinality u ; for, if F is a well-ordered set of cardinality u , then any bijection $F \rightarrow E$ transports the well-ordering of F to a well-ordering R of E , whence $F \approx E_R$ via the same bijection.

Let $\Lambda = \{\text{ord } E_R : R \in \mathcal{R}\}$ (a well-ordered set, by 1.14.24) and let α be the smallest element of Λ . Say $\alpha = \text{ord } E_S$, where $S \in \mathcal{R}$; let us show that α meets the requirements of the theorem.

Assuming β is an ordinal with $\text{card } \beta = u$, we must show that $\alpha \leq \beta$. Say $\beta = \text{ord } F$. As shown above, $F \approx E_R$ for some $R \in \mathcal{R}$, therefore $\beta = \text{ord } E_R \geq \text{ord } E_S = \alpha$. \diamond

1.14.33. *Definition.* With notations as in the preceding theorem, we call α the **initial ordinal** associated with the cardinal number u and we write $\alpha = \text{inord } u$.

For example, $\text{inord } \aleph_0 = \omega$ by 1.14.30, (v); by the same token, $\omega + 1$ is not of the form $\text{inord } u$ for any cardinal u (the only candidate for u is \aleph_0 , and it doesn't work).

1.14.34. *Remark.* Combining the preceding two definitions, we see that

$$\text{card}(\text{inord } u) = u$$

for every cardinal number u , whereas

$$\text{inord}(\text{card } \alpha) \leq \alpha$$

for every ordinal number α (with equality only when α is an initial ordinal).

\aleph_0 is the first infinite cardinal and $\omega = \text{inord } \aleph_0$ (which might well be denoted ω_0) is the first infinite ordinal. Let us take the next step:

1.14.35. Theorem. *There exists a smallest uncountable cardinal number (it is denoted \aleph_1).*

Proof. Consider $c = 2^{\aleph_0} = \text{card } \mathcal{P}(\mathbb{N})$ (see 1.13.15), let

$$S = \{\text{card } A : A \subset \mathcal{P}(\mathbb{N}), A \text{ uncountable}\},$$

(for example, $c \in S$) and let \aleph_1 be the smallest element of S (1.14.25); say $\aleph_1 = \text{ord } B$, where $B \subset \mathcal{P}(\mathbb{N})$ is uncountable. It will suffice to show that if u is any uncountable cardinal, then $\aleph_1 \leq u$. Assume to the contrary that $u < \aleph_1$ (1.12.20). Say $u = \text{card } E$. We know from $u < \aleph_1$ that $E \sim A$ for some $A \subset B$. Since E is uncountable, so is A , therefore $\text{card } A \in S$ and $\aleph_1 \leq \text{card } A = \text{card } E = u$, contrary to $u < \aleph_1$. \diamond

Obviously $\aleph_1 \leq 2^{\aleph_0}$; the Continuum Hypothesis is the proposition that $\aleph_1 = 2^{\aleph_0}$. In any case, $\text{inord } \aleph_1$ is clearly the smallest uncountable ordinal; we have arrived at "big omega" (also denoted ω_1):

1.14.36. *Definition.* $\Omega = \text{inord } \aleph_1$.

This train goes on forever, but this is where we get off.

Exercises

1. Let E be a well-ordered set, $f: E \rightarrow E$ an order morphism. If there exists at least one point $x \in E$ such that $f(x) \leq x$, then f has a fixed point (that is, $f(z) = z$ for some $z \in E$).

{Hint: Let $T = \{x \in E: f(x) \leq x\}$, note that $f(T) \subset T$ and consider the first element of T .}

2. If E is a well-ordered set, then every nonempty subset of E that is bounded above has a least upper bound.

3. $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ for all ordinal numbers α, β, γ .

4. Let α, β, γ be ordinal numbers.

(i) $\alpha < \beta \Leftrightarrow \beta = \alpha + \delta$ for some $\delta \neq 0$.

(ii) $\alpha + 1 = \beta + 1 \Rightarrow \alpha = \beta$.

(iii) For every finite ordinal n , $\alpha + n = \beta + n \Rightarrow \alpha = \beta$.

(iv) α is infinite $\Leftrightarrow \alpha = \omega + \delta$ for some δ .

(v) α is infinite $\Leftrightarrow n + \alpha = \alpha$ for every finite ordinal n .

(vi) For every finite ordinal n , $n + \alpha = n + \beta \Rightarrow \alpha = \beta$.

(vii) Left cancellation holds: $\gamma + \alpha = \gamma + \beta \Rightarrow \alpha = \beta$.

(viii) Right cancellation fails: $1 + \omega = 2 + \omega$. Also, if $\alpha = \omega$ and $\beta = \omega + 1$, then $\alpha + \omega = \beta + \omega$ but $\alpha \neq \beta$.

(ix) $\alpha = \alpha + \beta \Leftrightarrow \beta = 0$.

(x) If α is infinite and β is finite, then $\beta + \alpha = \alpha$.

(xi) If β is an ordinal number and $\alpha = \beta + \beta + \beta + \dots$ is defined in the way suggested by 1.14.29, then $\beta + \alpha = \alpha$.

{Hint: Say $\beta = \text{ord } F$. Let $E = \mathbb{N} \times F$, equipped with the lexicographic ordering (§1.11, Exercise 3), so that E is the union of the sequence of pairwise disjoint well-ordered sets $F_n = \{n\} \times F$ ($n = 0, 1, 2, 3, \dots$), with every element of F_n less than every element of $F_{n+1} \cup F_{n+2} \cup \dots$. If $G = F_1 \cup F_2 \cup \dots$, then $E = F_0 \cup G$ and $E \approx G$.}

5. An ordinal number $\alpha \neq 0$ is called a *limit ordinal* if it has no immediate predecessor, that is, α is not of the form $\beta + 1$ for any ordinal β .

(i) If $\alpha = \text{ord } E \neq 0$, then α is a limit ordinal if and only if E has no last element.

(ii) Every infinite initial ordinal is a limit ordinal.

(iii) $\omega + 1$ is infinite, but not a limit ordinal.

(iv) $\omega + \omega$ is a limit ordinal but is not initial.

(v) Every finite ordinal is an initial ordinal but is not a limit ordinal. Every limit ordinal is infinite.

6. Define \aleph_2 and ω_2 .

7. Let R be the usual ordering of \mathbb{N} and let R' be the reverse ordering (1.7.4); equipped with R' , \mathbb{N} has no smallest element hence is not well-ordered.

8. Let $T = \mathbb{N}(1) = \{0, 1\}$ ("T for 2") and consider the product set $T \times \mathbb{N} = \{(m, n) : m = 0, 1; n = 0, 1, 2, 3, \dots\}$. Abbreviating $(0, n)$ and $(1, n)$ to n and n' , respectively, the graph of $T \times \mathbb{N}$ takes the form of a $2 \times \infty$ matrix

$$\begin{pmatrix} 0 & 1 & 2 & 3 & \dots \\ 0' & 1' & 2' & 3' & \dots \end{pmatrix}.$$

(i) If $T \times \mathbb{N}$ is given the lexicographic ordering, then rows have priority over columns: everything in row 1 comes before everything in row 2 (within a row, it is business as usual); written on one line in increasing order, $T \times \mathbb{N}$ is

$$0, 1, 2, 3, \dots; 0', 1', 2', 3', \dots.$$

(ii) If instead columns are given priority—order from top to bottom in each column, but the elements in a particular column come before every element occurring in a column to its right—then the elements of $T \times \mathbb{N}$ written in increasing order are

$$0, 0', 1, 1', 2, 2', 3, 3', \dots.$$

This is called the *backward lexicographic order* (or 'reverse lexicographic order', or 'antilexicographic order').

{CAUTION: "Reverse lexicographic order" is not the reverse of the lexicographic order (cf. Exercise 7).}

(iii) The backward lexicographic order on $T \times \mathbb{N}$ can be described as follows: give $\mathbb{N} \times T$ the lexicographic order, then transport it to $T \times \mathbb{N}$ via the bijection $(n, m) \mapsto (m, n)$. It is a well-ordering of $T \times \mathbb{N}$ (cf. §1.11, Exercise 3).

(iv) If $T \times \mathbb{N}$ has the lexicographic order, then $\text{ord}(T \times \mathbb{N}) = \omega + \omega$.

(v) If $T \times \mathbb{N}$ is equipped with the backward lexicographic order, then its ordinal number is ω .

(vi) More generally, if E and F are well-ordered sets, let us write $(E \times F)'$ for the set $E \times F$ equipped with the backward lexicographic order (defined, as in (ii), by giving priority to the second coordinate). As indicated in (iii), $(E \times F)' \approx F \times E$ (where $F \times E$ is equipped with the lexicographic order); in particular, $(E \times F)'$ is well-ordered and $\text{ord}(E \times F)' = \text{ord}(F \times E)$.

9. Given ordinal numbers α and β , the *product* $\alpha\beta$ of α and β is the ordinal number defined as follows. Let $\alpha = \text{ord } E$, $\beta = \text{ord } F$, and define $\alpha\beta$ to be the ordinal number of the set $E \times F$ equipped with the backward lexicographic order; that is, in the notations of Exercise 8,

$$\alpha\beta = \text{ord}(E \times F)' = \text{ord}(F \times E).$$

In particular, $0\alpha = \alpha 0 = 0$.

(i) $2\omega = \omega$ and $\omega 2 = \omega + \omega$.

(ii) $\alpha(\beta\gamma) = (\alpha\beta)\gamma$ for all ordinals α, β, γ .

(iii) $\gamma(\alpha + \beta) = \gamma\alpha + \gamma\beta$ for all ordinals α, β, γ .

(iv) $(1 + 1)\omega = \omega \neq \omega + \omega$. {The price we paid for the distributive law (iii) was to order $E \times F$ backwards; a penalty is the failure of the other distributive law.}

{Hint: (i) Cf. Exercise 8, (iv) and (v).}

(ii) If E, F, G are well-ordered sets with $\alpha = \text{ord } E$, $\beta = \text{ord } F$, $\gamma = \text{ord } G$, then

$$\alpha(\beta\gamma) = (\text{ord } E)[\text{ord}(F \times G)'] = (\text{ord } E)[\text{ord}(G \times F)] = \text{ord}[(G \times F) \times E];$$

calculate $(\alpha\beta)\gamma$ similarly and cite §1.7, Exercise 4, (vi).

(iii) Suppose $\alpha = \text{ord } E$, $\beta = \text{ord } F$, $\gamma = \text{ord } G$ and $E \cap F = \emptyset$. Using the technique of part (ii), it all comes down to showing that

$$(E \cup F) \times G \approx (E \times G) \cup (F \times G),$$

where all products are ordered lexicographically and disjoint unions are ordered so that each element of the first term is less than each element of the second (cf. 1.14.29).}

10. If β is any ordinal and $\alpha = \beta\omega$, then $\beta + \alpha = \alpha$.

{Hint: This is part (xi) of Exercise 4 in disguise, but we now have a simple algebraic proof.}

11. Theorem 1.14.1 fails if “well-ordered” is replaced by “simply ordered”.

12. If Λ is a countable set of countable ordinals, then the ordinal β constructed in Theorem 1.14.26 is countable.

{Hint: §1.13, Exercise 1.}

1.15. Extended Real Numbers

The symbols $\pm\infty$ are essentially a way of dealing with unbounded sets of real numbers:

1.15.1. *Definition.* Let $-\infty$ and $+\infty$ be two symbols not in \mathbb{R} and write

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}.$$

{For example, $-\infty = \emptyset$ and $+\infty = \mathbb{R}$ will do just fine, as will $-\infty = *$ and $+\infty = \bullet$.} The usual ordering of \mathbb{R} is extended to $\overline{\mathbb{R}}$ by specifying that

$$-\infty < x \text{ and } x < +\infty \text{ for every } x \in \mathbb{R},$$

and that $-\infty < +\infty$; defining $a \leq b$ in $\overline{\mathbb{R}}$ to mean that $a < b$ or $a = b$, we have a simple ordering of $\overline{\mathbb{R}}$. The set $\overline{\mathbb{R}}$, equipped with this ordering, is called the set of **extended real numbers**. The real numbers $x \in \mathbb{R}$ are called the *finite* elements of $\overline{\mathbb{R}}$, $\pm\infty$ its *infinite* elements. (The rationale for the notation $\overline{\mathbb{R}}$ is given in 3.3.17 below.)

The usual interval notations, when extended to $\overline{\mathbb{R}}$, yield formulas such as

$$\mathbb{R} = (-\infty, +\infty),$$

$$\overline{\mathbb{R}} = [-\infty, +\infty],$$

$$(-\infty, 5) = \{x \in \mathbb{R} : x < 5\},$$

$$[0, +\infty] = \{x \in \mathbb{R} : x \geq 0\} \cup \{+\infty\}.$$

1.15.2. *Theorem.* Every nonempty subset of $\overline{\mathbb{R}}$ has a least upper bound (i.e., a smallest majorant) and a greatest lower bound (i.e., a largest minorant).

Proof. Let A be a nonempty subset of $\overline{\mathbb{R}}$. If $A \subset \{-\infty, +\infty\}$, the assertions are obvious. Otherwise, $A \cap \mathbb{R} \neq \emptyset$. If $+\infty \in A$ or if $A \cap \mathbb{R}$ is not bounded above in \mathbb{R} , then $+\infty$ is the smallest majorant of A in $\overline{\mathbb{R}}$; on the other hand, if $+\infty \notin A$ and $A \cap \mathbb{R}$ is bounded above in \mathbb{R} , then the least upper bound of $A \cap \mathbb{R}$ in \mathbb{R} (§1.8) is also the smallest majorant of A in $\overline{\mathbb{R}}$. The arguments for greatest lower bound are similar. \diamond

1.15.3. *Definition.* If A is a nonempty subset of $\overline{\mathbb{R}}$, its least upper bound (also called its **supremum**) is denoted $\sup A$, and its greatest lower bound (also called its **infimum**) is denoted $\inf A$.

1.15.4. *Algebraic operations in $\overline{\mathbb{R}}$.* The algebraic operations of \mathbb{R} can, at least partially, be extended to $\overline{\mathbb{R}}$:

(i) The operation $x \mapsto -x$ in \mathbb{R} is extended to $\overline{\mathbb{R}}$ by defining $-(+\infty) = -\infty$ and $-(-\infty) = +\infty$. For extended real numbers x and y , $x < y \Leftrightarrow -x > -y$.

(ii) The operation $x \mapsto |x|$ in \mathbb{R} is extended to $\overline{\mathbb{R}}$ by defining $|+\infty| = |-\infty| = +\infty$; concisely, $|\pm\infty| = +\infty$.

(iii) The product xy for elements of \mathbb{R} is extended to $\overline{\mathbb{R}}$ by the following conventions (guided in part by the theory of limits in elementary calculus):

$$\text{for } x > 0, \quad x(+\infty) = (+\infty)x = +\infty, \text{ and}$$

$$x(-\infty) = (-\infty)x = -\infty;$$

$$\text{for } x < 0, \quad x(+\infty) = (+\infty)x = -\infty, \text{ and}$$

$$x(-\infty) = (-\infty)x = +\infty;$$

$$\text{and } 0(\pm\infty) = (\pm\infty)0 = 0.$$

Then $xy = yx$ for all extended real numbers x and y .

(iv) The sum $x + y$ for elements of \mathbb{R} is *partially* extended to $\overline{\mathbb{R}}$ by the following conventions:

$$\text{for } x \in \mathbb{R} \text{ or } x = +\infty, \quad x + (+\infty) = (+\infty) + x = +\infty;$$

$$\text{for } x \in \mathbb{R} \text{ or } x = -\infty, \quad x + (-\infty) = (-\infty) + x = -\infty.$$

The sums $(+\infty) + (-\infty)$ and $(-\infty) + (+\infty)$ are **not defined**.

For extended real numbers x, y, z , each of the equations $x + y = y + x$ and $(x + y) + z = x + (y + z)$ holds whenever one of its sides is defined.

(v) For $x, y \in \overline{\mathbb{R}}$, one writes $x - y = x + (-y)$ provided the sum on the right side exists in the sense of (iv). The only differences that are **not** defined are $(+\infty) - (+\infty)$ and $(-\infty) - (-\infty)$.

The operations $x + y$ and xy are defined for all $x, y \in [0, +\infty]$, and both operations are commutative and associative. In fact, 'infinite sums' also exist in $[0, +\infty]$:

1.15.5. *Definition.* Let $(a_i)_{i \in I}$ be a family of elements of $[0, +\infty]$. For each finite subset F of I , write

$$a_F = \sum_{i \in F} a_i,$$

with the convention that $a_\emptyset = 0$; the **sum** of the family is defined by the formula

$$\sum_{i \in I} a_i = \sup \{ a_F : F \subset I, \text{ } F \text{ finite} \},$$

also written $\sum a_i$ when the index set is not in doubt.

1.15.6. *Remarks.* With the preceding notations, if F and G are finite subsets of I with $F \cap G = \emptyset$, then

$$a_{F \cup G} = a_F + a_G;$$

it follows that the mapping $F \mapsto a_F$ is increasing in the sense that $F \subset G \Rightarrow a_F \leq a_G$.

1.15.7. Theorem. If $(a_i)_{i \in I}$ and $(b_i)_{i \in I}$ are two families in $[0, +\infty]$ indexed by the same set I , then

$$\sum_{i \in I} (a_i + b_i) = \sum_{i \in I} a_i + \sum_{i \in I} b_i.$$

Proof. Write $c_i = a_i + b_i$, $c = \sum c_i$, $a = \sum a_i$, $b = \sum b_i$. For every finite set $F \subset I$,

$$c_F = a_F + b_F \leq a + b,$$

whence $c \leq a + b$.

It remains to show that $a + b \leq c$. If $c = +\infty$, this is trivial. Assuming $c < +\infty$, we have also $a < +\infty$ and $b < +\infty$. Fix a finite set $F \subset I$; it will suffice to show that $a_F \leq c - b$, that is, $b \leq c - a_F$. Fix another finite set $G \subset I$; it will suffice to show that $b_G \leq c - a_F$. Indeed,

$$a_F + b_G \leq a_{F \cup G} + b_{F \cup G} = c_{F \cup G} \leq c,$$

whence the desired inequality. \diamond

1.15.8. Corollary. If $(a_{ij})_{(i,j) \in I \times J}$ is a doubly indexed family of elements of $[0, +\infty]$, then

$$\sum_{(i,j) \in I \times J} a_{ij} = \sum_{i \in I} \left(\sum_{j \in J} a_{ij} \right) = \sum_{j \in J} \left(\sum_{i \in I} a_{ij} \right).$$

Proof. Write a for the sum on the left, b for the iterated sum in the middle, c for the iterated sum on the right. Let us show, for example, that $a = b$ (the proof that $a = c$ is similar). Writing

$$b_i = \sum_{j \in J} a_{ij}$$

for each $i \in I$, we have

$$b = \sum_{i \in I} b_i.$$

Proof that $a \leq b$: Given any finite set $H \subset I \times J$, it suffices to show that $a_H \leq b$. We have $H \subset F \times G$ for suitable finite sets $F \subset I$ and $G \subset J$, therefore

$$a_H \leq a_{F \times G} = \sum_{(i,j) \in F \times G} a_{ij} = \sum_{i \in F} \left(\sum_{j \in G} a_{ij} \right) \leq \sum_{i \in F} b_i = b_F \leq b.$$

Proof that $b \leq a$: Given any finite $F \subset I$, it suffices to show that $b_F \leq a$. Now,

$$b_F = \sum_{i \in F} b_i = \sum_{i \in F} \left(\sum_{j \in J} a_{ij} \right) = \sum_{j \in J} \left(\sum_{i \in F} a_{ij} \right)$$

(the last equality by the preceding theorem); thus, given any finite subset $G \subset J$, it suffices to show that

$$\sum_{j \in G} \left(\sum_{i \in F} a_{ij} \right) \leq a,$$

and this is true because the iterated sum on the left is $a_{F \times G}$. \diamond

If $a_i \in [0, +\infty]$ and $\sum a_i < +\infty$, we are essentially in the framework of positive-term infinite series (or possibly finite sums):

1.15.9. Theorem. If $(a_i)_{i \in I}$ is a family in $[0, +\infty]$ such that

$$\sum_{i \in I} a_i < +\infty,$$

then every a_i is finite and $a_i = 0$ for all but countably many i .

Proof. Let $a = \sum a_i$; for every i , $0 \leq a_i = a_{\{i\}} \leq a < \infty$, so a_i is finite. The claim is that the set $\{i \in I : a_i > 0\}$ is countable. Writing $I_n = \{i \in I : a_i \geq 1/n\}$ ($n = 1, 2, 3, \dots$), we have

$$\{i \in I : a_i > 0\} = \bigcup_{n=1}^{\infty} I_n,$$

so it will suffice to show that every I_n is finite. Fix a positive integer n . If $F \subset I_n$ is finite, say F has k elements, then $a \geq a_F \geq k(1/n)$, thus $k \leq na$; this shows that the number of elements of I_n is bounded by na . \diamond

With $\sum a_i < +\infty$ as in the preceding theorem, nothing essential is lost by replacing the index set I by the set $\{i \in I : a_i > 0\}$. For the denumerable index set $\mathbb{P} = \{1, 2, 3, \dots\}$ an alternate notation is available:

1.15.10. Definition. If $(a_n)_{n \in \mathbb{P}}$ is a sequence in $[0, +\infty]$, one defines

$$\sum_{k=1}^{\infty} a_k = \sup \left\{ \sum_{k=1}^n a_k : n \in \mathbb{P} \right\};$$

thus, writing

$$a_{\{1, \dots, n\}} = \sum_{k=1}^n a_k$$

as in Definition 1.15.5, we have

$$\sum_{k=1}^{\infty} a_k = \sup\{a_{\{1, \dots, n\}} : n \in \mathbb{P}\}.$$

1.15.11. Theorem. If $(a_n)_{n \in \mathbb{P}}$ is a sequence in $[0, +\infty]$, then

$$\sum_{n \in \mathbb{P}} a_n = \sum_{k=1}^{\infty} a_k.$$

Proof. Write a and b for the left and right sides, respectively. For every positive integer n ,

$$\sum_{k=1}^n a_k = a_{\{1, \dots, n\}} \leq a,$$

therefore $b \leq a$. On the other hand, for every finite set $F \subset \mathbb{P}$, there exists a positive integer n such that $F \subset \{1, \dots, n\}$, thus

$$a_F \leq a_{\{1, \dots, n\}} \leq b,$$

whence $a \leq b$. \diamond

Exercises

1. $|xy| = |x||y|$ for all extended real numbers x and y .
2. Let x, y, z be extended real numbers.
 - (i) If $x(y+z)$ and $xy+xz$ are both defined, then $x(y+z) = xy+xz$.
 - (ii) If $x = +\infty$, $y = -1$ and $z = +\infty$, then $x(y+z)$ is defined but $xy+xz$ is not.
 - (iii) If $x = 0$, $y = +\infty$ and $z = -\infty$, then $xy+xz$ is defined but $x(y+z)$ is not.
3. A partially ordered set I is said to be *upward directed* (or ‘upward filtering’) if, for every pair of elements i, j of I , there exists at least one $k \in I$ such that $i \leq k$ and $j \leq k$. Examples:
 - (i) Any simply ordered set.
 - (ii) Any partially ordered set with a largest element.
 - (iii) The set of all finite subsets of a given set X , with inclusion as the order relation.
4. Let X be a set. A *net* in X is a family $(x_i)_{i \in I}$ of elements of X indexed by a partially ordered set I that is upward directed (cf. Exercise 3). A net $(x_i)_{i \in I}$ in a partially ordered set X is said to be *increasing* if $i \leq j \Rightarrow x_i \leq x_j$ (in other words, $i \mapsto x_i$ is an order morphism $I \rightarrow X$), *decreasing* if $i \leq j \Rightarrow x_i \geq x_j$, and *monotone* if it is either increasing or

decreasing; the terms 'strictly increasing', 'strictly decreasing' and 'strictly monotone' are defined in the obvious way. We write $x_i \uparrow$ (or $x_i \downarrow$) to indicate nets that are increasing (or decreasing); if, in addition, the set $\{x_i : i \in I\}$ has a least upper bound (or a greatest lower bound) x in X , we write $x_i \uparrow x$ (or $x_i \downarrow x$), sometimes adding phrases 'as $i \uparrow$ ', 'as $i \uparrow \infty$ ', or 'as $i \uparrow$ in I ', as a reminder that an ordering of the index set I is involved.

(i) If, as in Definition 1.15.5,

$$a = \sum_{i \in I} a_i \quad \text{and} \quad a_F = \sum_{i \in F} a_i$$

for finite sets $F \subset I$, then $a_F \uparrow a$ as $F \uparrow$ (in the upward directed set of finite subsets of I , ordered by inclusion).

(ii) If $(x_i)_{i \in I}$ and $(y_i)_{i \in I}$ are increasing nets in $[0, +\infty]$ with same (upward directed) index set I , and if $x_i \uparrow x$ and $y_i \uparrow y$, then $x_i + y_i \uparrow x + y$ and $x_i y_i \uparrow xy$.

{Hint: Cf. the proof of 1.15.7. For the assertion about products, the convention $0 \cdot (+\infty) = 0$ plays a role.}

(iii) With notations as in (i), $c \sum a_i = \sum c a_i$ for every $c \in [0, +\infty]$.

{Hint: Apply (ii) to the situation of (i), with $x_i = c$ for all i .}

5. Let $(a_i)_{i \in I}$ be a family in $[0, +\infty]$, $\sigma : I \rightarrow I$ a mapping of the index set into itself, and let $b_i = a_{\sigma(i)}$ for all $i \in I$. Write $a = \sum_{i \in I} a_i$, $b = \sum_{i \in I} b_i$.

(i) If σ is injective, then $b \leq a$.

(ii) If σ is surjective, then $a \leq b$.

(iii) (Commutative law) If $\sigma : I \rightarrow I$ is bijective, then

$$\sum_{i \in I} a_i = \sum_{i \in I} a_{\sigma(i)}.$$

{Hint: (i) If σ is injective and F is a finite subset of I , then σ maps F bijectively onto $\sigma(F)$ and, in the notation of 1.15.5, $b_F = a_{\sigma(F)}$.

(ii) If σ is surjective and F is a finite subset of I , then there exists a finite subset G of I such that σ maps G bijectively onto F .}

6. If $(a_i)_{i \in I}$ is a family in $[0, +\infty]$, then (with notations as in 1.15.5),

$$a_{F \cup G} + a_{F \cap G} = a_F + a_G$$

for all finite subsets F and G of I .

{Hint: If either term on the right is $+\infty$, the assertion is obvious.}

7. Let (X, \leq) be a partially ordered set in which every nonempty subset has a supremum (for example, $X = \mathbb{R}$ or $X = [0, +\infty]$, ordered in the usual way).

(i) If $(x_i)_{i \in I}$ is a family of elements of X and if $(I_j)_{j \in J}$ is a family of nonempty subsets of I with $I = \bigcup_{j \in J} I_j$, then

$$\sup_{i \in I} x_i = \sup_{j \in J} \left(\sup_{i \in I_j} x_i \right).$$

(This principle is called the 'associativity of sups'.)

(ii) If $(J_i)_{i \in I}$ is a family of nonempty sets and if, for each $i \in I$, $(x_{ij})_{j \in J_i}$ is a family of elements of X , then

$$\sup_{i \in I} \left(\sup_{j \in J_i} x_{ij} \right) = \sup \{ x_{ij} : i \in I, j \in J_i \}.$$

(iii) If $(x_{ij})_{(i,j) \in I \times J}$ is a doubly-indexed family of elements of X , then

$$\sup_{i \in I} \left(\sup_{j \in J} x_{ij} \right) = \sup \{ x_{ij} : (i, j) \in I \times J \} = \sup_{j \in J} \left(\sup_{i \in I} x_{ij} \right).$$

1.16. limsup, liminf, Convergence in $\overline{\mathbb{R}}$

The following notations are fixed for the section: (a_n) is a sequence in the set $\overline{\mathbb{R}}$ of extended real numbers (1.15.1); for each positive integer n , we write

$$b_n = \sup \{ a_k : k \geq n \} = \sup_{k \geq n} a_k,$$

$$c_n = \inf \{ a_k : k \geq n \} = \inf_{k \geq n} a_k.$$

As n increases, the set $\{a_k : k \geq n\}$ can only shrink; more precisely, the correspondence $n \mapsto \{a_k : k \geq n\}$ is a decreasing function $\mathbb{P} \rightarrow \mathcal{P}(\overline{\mathbb{R}})$, where $\mathcal{P}(\overline{\mathbb{R}})$ is ordered by inclusion. It follows that

$$b_n \downarrow \text{ and } c_n \uparrow \text{ as } n \uparrow.$$

Let

$$b = \inf \{ b_n : n \in \mathbb{P} \} = \inf_{n \geq 1} b_n,$$

$$c = \sup \{ c_n : n \in \mathbb{P} \} = \sup_{n \geq 1} c_n.$$

We write $b_n \downarrow b$ to express the fact that (b_n) is a decreasing sequence with infimum b , and $c_n \uparrow c$ has the dual meaning (cf. §1.15, Exercise 4).

1.16.1. Definition. With the preceding notations, b is called the **limit superior** of the sequence (a_n) , written

$$\limsup a_n = \inf_{n \geq 1} \left(\sup_{k \geq n} a_k \right);$$

c is called the **limit inferior** of the sequence (a_n) , written

$$\liminf a_n = \sup_{n \geq 1} \left(\inf_{k \geq n} a_k \right).$$

Scholium. To calculate the \limsup of the sequence (a_n) , truncate the sequence at n and take \sup , then take the \inf over all possible truncations. Similarly for $\liminf a_n$: truncate at n and take \inf , then take \sup over all truncations.

1.16.2. Theorem. For every sequence (a_n) in $\overline{\mathbb{R}}$,

$$\liminf a_n \leq \limsup a_n.$$

Proof. Fix positive integers m and n ; we are going to show that $c_m \leq b_n$. Let $r = \max\{m, n\}$; since a_r belongs to both $\{a_k : k \geq m\}$ and $\{a_k : k \geq n\}$, we have $c_m \leq a_r \leq b_n$. Varying m , we have

$$c = \sup_{m \geq 1} c_m \leq b_n,$$

and, varying n , this yields

$$c \leq \inf_{n \geq 1} b_n = b. \diamond$$

1.16.3. Theorem. For every sequence (a_n) in $\overline{\mathbb{R}}$,

$$\liminf a_n = -\limsup(-a_n),$$

$$\limsup a_n = -\liminf(-a_n).$$

Proof. This follows from the fact that $x \mapsto -x$ is an order-reversing bijection $\overline{\mathbb{R}} \rightarrow \overline{\mathbb{R}}$. Explicitly, write $a'_n = -a_n$ and let b'_n, b', c'_n, c' be the quantities computed from (a'_n) analogous to the quantities b_n, b, c_n, c computed from (a_n) . Since

$$\{a'_k : k \geq n\} = \{-a_k : k \geq n\}$$

we have $b'_n = -c_n$, whence $b' = -c$; this is the first of the desired formulas, and the second follows from it by the substitution $a_n \mapsto -a_n$. \diamond

A useful characterization of the limit superior:

1.16.4. Theorem. Let (a_n) be a sequence in $\overline{\mathbb{R}}$ and let $b = \limsup a_n$. Then

$$(1) \ r > b \Rightarrow a_n < r \text{ ultimately};$$

$$(2) \ r < b \Rightarrow a_n > r \text{ frequently}.$$

Moreover, these two properties determine b uniquely.

Proof. "Ultimately" means "from some index onward"; "frequently" means "for infinitely many indices".

Proof of (1): Let r be an extended real number with $r > b$ (if $b = +\infty$ then no such r exists and the implication is vacuously true); we seek an index m such that $a_k < r$ for all $k \geq m$. Since $r > b$ and b is the largest minorant of the set $B = \{b_n : n \in \mathbb{P}\}$, r is not a minorant of B ; therefore $r > b_m$ for some m . Since $b_m = \sup\{a_k : k \geq m\}$, we have $k \geq m \Rightarrow a_k \leq b_m < r$, whence the assertion (1).

Proof of (2): The negation of " $a_n > r$ frequently" is " $a_n \leq r$ ultimately"; arguing contrapositively, let us show that

$$a_n \leq r \text{ ultimately} \Rightarrow b \leq r.$$

Assuming $a_k \leq r$ for all $k \geq m$, we have $b \leq b_m = \sup\{a_k : k \geq m\} \leq r$.

Finally, suppose that $b' \in \overline{\mathbb{R}}$ satisfies the conditions (1') and (2') analogous to (1) and (2); the claim is that $b' = b$. Assume to the contrary, for example, that $b < b'$. Choose r with $b < r < b'$. Since $r > b$, we have $a_n < r$ ultimately by (1), and since $r < b'$ we have $a_n > r$ frequently by (2'); these two statements are contradictory. \diamond

Arguing dually (or combining the preceding theorem with 1.16.3) we have

1.16.5. Theorem. Let (a_n) be a sequence in $\overline{\mathbb{R}}$, and let $c = \liminf a_n$. Then:

(3) $r < c \Rightarrow a_n > r$ ultimately;

(4) $r > c \Rightarrow a_n < r$ frequently.

These two properties determine c uniquely.

Always $\liminf a_n \leq \limsup a_n$; when do we have equality? Theorems 1.16.4 and 1.16.5 yield the following criterion:

1.16.6. Theorem. For a sequence (a_n) in $\overline{\mathbb{R}}$, the following conditions are equivalent:

(a) $\liminf a_n = \limsup a_n$;

(b) there exists an extended real number a such that

(i) $r > a \Rightarrow a_n < r$ ultimately, and

(ii) $r < a \Rightarrow a_n > r$ ultimately.

When the conditions hold,

$$a = \liminf a_n = \limsup a_n$$

(in particular, a is unique).

Proof. (a) \Rightarrow (b): We are assuming that $c = b$; write $a = c = b$. Condition (i) is (1) of 1.16.4, and condition (ii) is (3) of 1.16.5.

(b) \Rightarrow (a): Assuming $a \in \overline{\mathbb{R}}$ satisfies (i) and (ii), we must show that $a = b$ and $a = c$. To prove, for example, that $a = b$, it suffices to verify that a satisfies (1) and (2) of 1.16.4, and these are immediate from (i) and (ii). \diamond

1.16.7. Corollary. If (a_n) is a bounded sequence in \mathbb{R} , then

$$\liminf a_n = \limsup a_n \Leftrightarrow (a_n) \text{ is convergent in } \mathbb{R}.$$

When this is the case,

$$\lim_{n \rightarrow \infty} a_n = \liminf a_n = \limsup a_n.$$

Proof. Say $|a_n| \leq M$ for all n , where $0 \leq M < +\infty$; then b_n , c_n , b and c all belong to $[-M, M]$. Since (b_n) and (c_n) are bounded monotone sequences, we have $b_n \rightarrow b$ and $c_n \rightarrow c$ (convergence in \mathbb{R}).

\Rightarrow : By assumption, $c = b$; write $a = c = b$. For every positive real number ϵ , $a - \epsilon < a < a + \epsilon$; by condition (b) of the theorem, ultimately $a_n - \epsilon < a < a_n + \epsilon$, that is, $|a_n - a| < \epsilon$. This shows that $a_n \rightarrow a$ in \mathbb{R} .

\Leftarrow : Suppose there exists $a \in \mathbb{R}$ such that $a_n \rightarrow a$. Let $\epsilon > 0$ and choose an index N such that, for every $n \geq N$, $|a_n - a| \leq \epsilon$, that is, $a - \epsilon \leq a_n \leq a + \epsilon$. It follows that

$$n \geq N \Rightarrow a - \epsilon \leq c_n \leq b_n \leq a + \epsilon;$$

letting $n \rightarrow \infty$, we have $a - \epsilon \leq c \leq b \leq a + \epsilon$, and since this is true for every $\epsilon > 0$, we conclude that $a = c = b$. \diamond

The corollary points the way to a definition of convergence in $\overline{\mathbb{R}}$:

1.16.8. *Definition.* A sequence (a_n) in $\overline{\mathbb{R}}$ is said to be **convergent** if $\liminf a_n = \limsup a_n$; the common value of the \liminf and \limsup is called the **limit** of the sequence (a_n) , written

$$\lim_{n \rightarrow \infty} a_n$$

(or briefly $\lim a_n$). If $a = \lim a_n$, the sequence (a_n) is said to **converge** to a , and one writes

$$a_n \rightarrow a \text{ as } n \rightarrow \infty$$

(or simply $a_n \rightarrow a$).

1.16.9. *Remarks.* It is instructive to analyze the criteria (i), (ii) of 1.16.6 for convergence $a_n \rightarrow a$, according to the value of a .

case 1: $a = +\infty$.

(i) is vacuous. (ii) says that if $r < +\infty$ then $a_n > r$ ultimately; equivalently,

$$K \in \mathbb{R}, K > 0 \Rightarrow a_n > K \text{ ultimately}$$

(and one need only consider positive integral values of K).

case 2: $a = -\infty$.

(ii) is vacuous. (i) says that if $r > -\infty$ then $a_n < r$ ultimately; equivalently,

$$K \in \mathbb{R}, K > 0 \Rightarrow a_n < -K \text{ ultimately.}$$

case 3: $a \in \mathbb{R}$.

Then $a - 1 < a < a + 1$, so $a - 1 < a_n < a + 1$ ultimately. Thus, apart from a possible finite number of terms equal to $\pm\infty$, the sequence (a_n) is bounded. Dropping finitely many terms changes at most finitely many b_n and c_n , and b, c not at all; this case is essentially “business as usual” for a convergent sequence in \mathbb{R} .

1.16.10. *Example.* Define a function $f : [-1, 1] \rightarrow \overline{\mathbb{R}}$ by the formulas $f(-1) = -\infty$, $f(1) = +\infty$ and

$$f(x) = \frac{x}{1 - x^2} \quad \text{for } |x| < 1.$$

Since f is bijective and order-preserving, for every sequence (x_n) in $[-1, 1]$ we have

$$\begin{aligned} f(\liminf x_n) &= \liminf f(x_n), \\ f(\limsup x_n) &= \limsup f(x_n); \end{aligned}$$

it then follows from the definition of convergence in $\overline{\mathbb{R}}$ that $x_n \rightarrow x \Leftrightarrow f(x_n) \rightarrow f(x)$.

Convergence in $\overline{\mathbb{R}}$ was defined by means of \liminf and \limsup ; in turn, \liminf and \limsup can be analyzed in terms of convergence:

1.16.11. *Theorem.* Let (a_n) be a sequence in $\overline{\mathbb{R}}$, $c = \liminf a_n$, $b = \limsup a_n$, and let

$$A = \{x \in \overline{\mathbb{R}} : a_{n_k} \rightarrow x \text{ for a subsequence } (a_{n_k}) \text{ of } (a_n)\}.$$

Then

$$\{c, b\} \subset A \subset [c, b],$$

thus b is the largest element of A , and c is the smallest. In particular, every sequence in $\overline{\mathbb{R}}$ has at least one convergent subsequence.

Proof. For the first inclusion let us show, for example, that $b \in A$.

case 1: $b = +\infty$.

Then $b > 1$, so $a_n > 1$ frequently by (2) of 1.16.4. Choose n_1 with $a_{n_1} > 1$. But also $b > 2$, so $a_n > 2$ frequently; choose $n_2 > n_1$ so that $a_{n_2} > 2$. Continuing, we obtain a subsequence (a_{n_k}) of (a_n) such that $a_{n_k} > k$ for all k , whence $a_{n_k} \rightarrow +\infty$; in particular, $b = +\infty \in A$.

case 2: $b = -\infty$.

Then $-\infty \leq c \leq b = -\infty$, so $c = b = -\infty$ and $a_n \rightarrow -\infty$; in particular, $b = -\infty \in A$.

case 3: $b \in \mathbb{R}$.

For every $\epsilon > 0$, it is clear from 1.16.4 that $b - \epsilon < a_n < b + \epsilon$ frequently; that is, for infinitely many n , a_n is finite and $|a_n - b| < \epsilon$. Let $\epsilon = 1$ and choose n_1 so that $|a_{n_1} - b| < 1$. Let $\epsilon = 1/2$; since $|a_n - b| < 1/2$

frequently, we can choose $n_2 > n_1$ so that $|a_{n_2} - b| < 1/2$. Continuing, we obtain a subsequence (a_{n_k}) of (a_n) such that $|a_{n_k} - b| < 1/k$ for all k , whence $a_{n_k} \rightarrow b$ and $b \in A$.

Similarly, $c \in A$ (alternatively, use 1.16.3). Thus $\{c, b\} \subset A$.

Finally, given $x \in A$, we must show that $c \leq x \leq b$. Let us show, for example, that $x \leq b$. Assume to the contrary that $x > b$, and choose r so that $b < r < x$. Since $b_n \downarrow b$ and $b < r$, there exists an index m such that $b_m < r$. Then

$$n \geq m \Rightarrow a_n \leq b_m < r,$$

thus $a_n < r$ ultimately. By hypothesis, $x \in A$, thus $a_{n_k} \rightarrow x$ for a suitable subsequence of (a_n) ; since $r < x$, it follows that $a_{n_k} > r$ for infinitely many k , contrary to the fact that $a_n < r$ ultimately. \diamond

Note that the last assertion of the theorem also follows from Example 1.16.10 and the Weierstrass-Bolzano theorem; indeed, the entire theorem follows at once from 1.16.10 and the corresponding theorem for bounded sequences in \mathbb{R} .¹

Exercises

1. With notations as in Theorem 1.16.11, (a_n) is convergent if and only if $A = \{a\}$ for some $a \in \overline{\mathbb{R}}$.

2. In the notations of Theorem 1.16.4, the condition $r = b$ is inconclusive. For example, if (a_n) is the sequence

$$-1, 0, -1/2, 0, -1/3, 0, \dots$$

then $b = 0$ but neither of the conditions " $a_n < 0$ ultimately" or " $a_n > 0$ frequently" holds.

3. Given a power series $(*) \sum_{k=0}^{\infty} a_k x^k$ with real (or complex) coefficients a_k , let $L = \limsup_{k \geq 1} |a_k|^{1/k}$. Prove:

(i) $L = 0 \Leftrightarrow |a_k|^{1/k} \rightarrow 0$.

(ii) $L = +\infty \Leftrightarrow |a_k|^{1/k}$ is unbounded.

(iii) If $x \neq 0$ and $1/|x| > L$ then $(*)$ is absolutely convergent.

(iv) If $x \neq 0$ and $1/|x| < L$ then $(*)$ is divergent.

(v) $L = 0 \Leftrightarrow (*)$ is absolutely convergent for every x (in \mathbb{R} or in \mathbb{C}).

(vi) $L = +\infty \Leftrightarrow (*)$ converges only for $x = 0$.

Let $R = 1/L$, with the convention that $1/+\infty = 0$ and $1/0 = +\infty$. Then:

(vii) $L = 1/R$.

¹ First course, p. 53, 3.7.7.

(viii) $|x| < R \Rightarrow (*)$ is absolutely convergent.

(ix) $|x| > R \Rightarrow (*)$ is divergent.

(x) $R = 0 \Leftrightarrow (*)$ converges only for $x = 0$.

(xi) $R = +\infty \Leftrightarrow (*)$ is absolutely convergent for every x (in \mathbb{R} or in \mathbb{C}).

{Hints: (iii) If $L < r < 1/|x|$ then $|a_k|^{1/k} < r$ ultimately; compare $(*)$ with the geometric series $\sum_{k=0}^{\infty} (rx)^n$. (iv) If $1/|x| < L$ then $|a_k|^{1/k} > 1/|x|$ frequently; infer that $a_k x^k \not\rightarrow 0$.}

CHAPTER 2

Lebesgue Measure

- §2.1. Lebesgue outer measure on \mathbb{R}
- §2.2. Measurable sets
- §2.3. Cantor set: an uncountable set of measure zero
- §2.4. Borel sets, regularity
- §2.5. A nonmeasurable set
- §2.6. Abstract measure spaces

One of the aims of the Lebesgue theory is to assign to each subset A of \mathbb{R} an element of $[0, +\infty]$, to be thought of as the ‘size’ of A , in such a way that the size of a bounded interval is its length, and the function

$$(*) \quad A \mapsto \text{size of } A$$

is well-behaved for as many sets A as possible. The facts are roughly as follows: (1) it is possible to assign a size to every subset of \mathbb{R} , but the function $(*)$ is not well-behaved; (2) on a large class of subsets of \mathbb{R} (including the intervals) the function $(*)$ is well-behaved. Thus, there is a trade-off between the extent of the domain of a size function and the goodness of its behavior.

More precisely, to each subset A of \mathbb{R} there is assigned a nonnegative extended real number $\lambda^*(A)$, called the *outer measure* of A ; certain subsets of \mathbb{R} are singled out and called *measurable*; the restriction of λ^* to the class \mathcal{M} of measurable sets is well-behaved and is called *Lebesgue measure*.

2.1. Lebesgue Outer Measure on \mathbb{R}

The point of departure is the concept of interval length:

2.1.1. *Definition.* If I is a bounded interval in \mathbb{R} , with endpoints a and b ($a \leq b$), the **length** of I is the nonnegative real number $\lambda(I) = b - a$.

Built into the definition is an indifference as to whether or not the endpoints a, b belong to I ; for example, $\lambda([a, b]) = \lambda((a, b))$. In effect, the

finite set $\{a, b\}$ (the set-theoretic gap between the two intervals) has been declared to be 'negligible'; the edifice of the Lebesgue theory rests on a precise definition of this word:

2.1.2. *Definition.* A set $A \subset \mathbb{R}$ is said to be **negligible** if it can be covered by a sequence of open intervals whose total length is arbitrarily small. More precisely, for every $\epsilon > 0$ there exists a sequence (I_n) of open intervals such that

$$A \subset \bigcup_{n=1}^{\infty} I_n \quad \text{and} \quad \sum_{n=1}^{\infty} \lambda(I_n) < \epsilon.$$

2.1.3. *Remarks.* (i) Since $\emptyset = (a, a)$, coverings by finitely many intervals are permitted.

(ii) The *kind* of interval (open, closed, semiclosed) is immaterial: in a covering by open intervals, including the endpoints does not change the lengths; in a covering by closed intervals $J_n = [a_n, b_n]$ of total length $< \epsilon$, the intervals can be enlarged to open intervals $I_n = (a_n - \delta_n, b_n + \delta_n)$ with total length $< 2\epsilon$, for example by choosing $\delta_n = \epsilon/2^{n+1}$.

(iii) Every countable set in \mathbb{R} is negligible. For, if $A = \{a_n : n \in \mathbb{P}\}$ then the (degenerate) intervals $J_n = [a_n, a_n]$ form a covering of A by a sequence of closed intervals of total length 0, therefore A is negligible by the preceding remark.

(iv) An example of an uncountable negligible set is given in §2.3.

(v) A striking application of the concept of negligibility is Lebesgue's criterion for Riemann-integrability: *A bounded function $f : [a, b] \rightarrow \mathbb{R}$ is Riemann-integrable if and only if its set of discontinuities is negligible.* This is proved in §5.13.

A set $A \subset \mathbb{R}$ is negligible if and only if the set of all sums $\sum \lambda(I_n)$, where (I_n) is a sequence of intervals with $A \subset \bigcup I_n$, has infimum 0. This suggests a way of defining the size of an arbitrary subset of \mathbb{R} :

2.1.4. *Definition.* For every subset A of \mathbb{R} , the (Lebesgue) **outer measure** of A , denoted $\lambda^*(A)$, is defined by the formula

$$\lambda^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \lambda(I_n) : A \subset \bigcup I_n \right\},$$

where (I_n) varies over all possible sequences of open intervals of \mathbb{R} whose union contains A .

2.1.5. *Remarks.* (i) $\lambda^*(A)$ is the infimum of a nonempty subset of $[0, +\infty]$; for example, this subset contains $+\infty$ since $A \subset \mathbb{R} = \bigcup (-n, n)$.

(ii) $0 \leq \lambda^*(A) \leq +\infty$ for every $A \subset \mathbb{R}$, thus $\lambda^* : \mathcal{P}(\mathbb{R}) \rightarrow [0, +\infty]$.

(iii) $A \subset \mathbb{R}$ is negligible $\Leftrightarrow \lambda^*(A) = 0$.

The basic formal properties of Lebesgue outer measure are gathered in the following theorem:

2.1.6. Theorem. If λ^* is Lebesgue outer measure (2.1.4) then

- (1) $0 \leq \lambda^*(A) \leq +\infty$ for all $A \subset \mathbb{R}$;
- (2) $\lambda^*(\emptyset) = 0$;
- (3) λ^* is a monotone increasing set function, in the sense that $A \subset B \Rightarrow \lambda^*(A) \leq \lambda^*(B)$;
- (4) λ^* is countably subadditive, in the sense that

$$\lambda^*\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \lambda^*(A_n)$$

for every sequence (A_n) of subsets of \mathbb{R} ;

- (5) $\lambda^*(I) = \lambda(I)$ for every bounded interval I of \mathbb{R} .

Proof. (1) Already noted in 2.1.5.

- (2) $\emptyset \subset (1, 1)$ shows that $0 \leq \lambda^*(\emptyset) \leq 1 - 1$.

(3) Suppose $A \subset B \subset \mathbb{R}$. If (I_n) is a sequence of open intervals with $B \subset \bigcup I_n$, then also $A \subset \bigcup I_n$, therefore $\lambda^*(A) \leq \sum \lambda(I_n)$; varying the covering (I_n) of B , $\lambda^*(A) \leq \lambda^*(B)$.

(4) If the sum on the right is $+\infty$, the inequality is trivial. Suppose $\sum \lambda^*(A_n) < +\infty$; then $\lambda^*(A_n) < +\infty$ for all n . Let $\epsilon > 0$. For each n , choose a sequence of open intervals I_{nk} ($k = 1, 2, 3, \dots$) such that

$$A_n \subset \bigcup_{k=1}^{\infty} I_{nk} \quad \text{and} \quad \sum_{k=1}^{\infty} \lambda(I_{nk}) < \lambda^*(A_n) + \epsilon/2^n$$

(possible because $\lambda^*(A_n)$ is defined as an infimum and is finite). Then

$$\bigcup_{n=1}^{\infty} A_n \subset \bigcup_{n,k} I_{nk},$$

therefore (see 1.15.8 for the maneuvers with infinite sums)

$$\begin{aligned} \lambda^*\left(\bigcup_{n=1}^{\infty} A_n\right) &\leq \sum_{n,k} \lambda(I_{nk}) = \sum_{n=1}^{\infty} \left(\sum_{k=1}^{\infty} \lambda(I_{nk})\right) \\ &< \sum_{n=1}^{\infty} [\lambda^*(A_n) + \epsilon/2^n] = \sum_{n=1}^{\infty} \lambda^*(A_n) + \epsilon, \end{aligned}$$

and (4) results on varying ϵ .

- (5) Let a and b be the endpoints of I , so that

$$(a, b) \subset I \subset [a, b] = (a, b) \cup \{a, b\};$$

by the properties of λ^* already noted,

$$\lambda^*((a, b)) \leq \lambda^*(I) \leq \lambda^*([a, b]) \leq \lambda^*((a, b)) + \lambda^*([a, b]) = \lambda^*((a, b)),$$

therefore

$$(i) \quad \lambda^*(I) = \lambda^*((a, b)) = \lambda^*([a, b]).$$

From $(a, b) \subset (a, b)$ we see that $\lambda^*((a, b)) \leq \lambda((a, b)) = b - a$, thus

$$(ii) \quad \lambda^*(I) \leq b - a = \lambda(I);$$

we need only show the reverse inequality. In view of (i), we can suppose that $I = [a, b]$; assuming

$$(*) \quad [a, b] \subset \bigcup_{n=1}^{\infty} (a_n, b_n),$$

we need only show that

$$b - a \leq \sum_{n=1}^{\infty} (b_n - a_n)$$

(the infimum of such sums being equal to $\lambda^*(I)$). By the Heine-Borel theorem, it follows from (*) that

$$[a, b] \subset \bigcup_{k=1}^n (a_k, b_k)$$

for some n ; it will suffice to infer that

$$(**) \quad b - a < \sum_{k=1}^n (b_k - a_k).$$

We prove (**) by induction on n . For $n = 1$ it is obvious. Let $n \geq 2$ and assume that all's well with $n - 1$. Reordering the (a_k, b_k) if necessary, we can suppose that $a \in (a_n, b_n)$, thus $a_n < a < b_n$.

case 1: $b_n > b$.

Then $a_n < a \leq b < b_n$ and (**) is obvious.

case 2: $b_n \leq b$.

Then $a_n < a < b_n \leq b$, so $[b_n, b]$ is disjoint from (a_n, b_n) ; but

$$[b_n, b] \subset [a, b] \subset \bigcup_{k=1}^n (a_k, b_k),$$

therefore

$$[b_n, b] \subset \bigcup_{k=1}^{n-1} (a_k, b_k).$$

By the induction hypothesis,

$$b - b_n < \sum_{k=1}^{n-1} (b_k - a_k),$$

which, added to $b_n - a < b_n - a_n$, yields (**). \diamond

2.1.7. Corollary. *Every subset of a negligible set is negligible, and a countable union of negligible sets is negligible.*

Proof. Immediate from (3) and (4) of the theorem. \diamond

2.1.8. Corollary. *If I is an unbounded interval, then $\lambda^*(I) = +\infty$.*

Proof. If, for example, $I = (-\infty, c]$ then, for every positive integer n , $I \supset [c - n, c]$, therefore $\lambda^*(I) \geq n$ by (3) and (5) of the theorem. \diamond

The definition of outer measure specified coverings by open intervals, but the type of interval is not critical:

2.1.9. Corollary. *For every subset A of \mathbb{R} ,*

$$\lambda^*(A) = \inf \left\{ \sum \lambda(J_n) : A \subset \bigcup J_n \right\},$$

where (J_n) runs over all countable coverings of A by bounded intervals that are closed (or left-closed and right-open, or right-closed and left-open, or a mixture of the four types).

Proof. Let $A \subset \mathbb{R}$ and write α for the infimum on the right side. If $A \subset \bigcup J_n$ is a covering of the indicated type, then (by the theorem)

$$\lambda^*(A) \leq \lambda^*\left(\bigcup J_n\right) \leq \sum \lambda^*(J_n) = \sum \lambda(J_n);$$

varying the covering, $\lambda^*(A) \leq \alpha$.

Suppose (I_n) is a sequence of open intervals with $A \subset \bigcup I_n$. Let J_n be the interval of the contemplated type (closed, etc.) having the same endpoints as I_n . Then $I_n \subset J_n$, therefore $A \subset \bigcup J_n$ and

$$\alpha \leq \sum \lambda(J_n) = \sum \lambda(I_n)$$

by the definition of α (and of λ); varying the covering (I_n) , we have $\alpha \leq \lambda^*(A)$. \diamond

A key property of Lebesgue outer measure is its invariance under translation:

2.1.10. Theorem. $\lambda^*(A + c) = \lambda^*(A)$ for all $A \subset \mathbb{R}$ and $c \in \mathbb{R}$.

Proof. The set $A + c = \{x + c : x \in A\}$ is the image of A under the order isomorphism $x \mapsto x + c$ of \mathbb{R} onto \mathbb{R} . For an open interval $I = (a, b)$, $I + c = (a + c, b + c)$ and $\lambda(I + c) = \lambda(I)$. If (I_n) is a sequence of open intervals with $A \subset \bigcup I_n$, then $A + c \subset \bigcup (I_n + c)$ and

$$\lambda^*(A + c) \leq \sum \lambda(I_n + c) = \sum \lambda(I_n);$$

varying the covering (I_n) , $\lambda^*(A + c) \leq \lambda^*(A)$. The inequality proves its own reverse: $\lambda^*(A) = \lambda^*((A + c) + (-c)) \leq \lambda^*(A + c)$, whence equality. \diamond

For multiplication by c , there is a factor of proportionality:

2.1.11. Theorem. $\lambda^*(cA) = |c| \cdot \lambda^*(A)$ for all $A \subset \mathbb{R}$ and $c \in \mathbb{R}$.

Proof. Here $cA = \{cx : x \in A\}$. If $A = \emptyset$ then $cA = \emptyset$ and both sides of the asserted equality are 0. Assume $A \neq \emptyset$.

case 1: $c = 0$.

Then $cA = \{0\}$, $\lambda^*(cA) = 0$ and $|c| \cdot \lambda^*(A) = 0$ (even if $\lambda^*(A) = +\infty$).

case 2: $c > 0$.

The argument is analogous to 2.1.10. Again $x \mapsto cx$ is an order isomorphism $\mathbb{R} \rightarrow \mathbb{R}$; if $I = (a, b)$ then $cI = (ca, cb)$ and $\lambda(cI) = c\lambda(I)$. If (I_n) is a sequence of open intervals with $A \subset \bigcup I_n$, then $cA \subset \bigcup cI_n$ and

$$\lambda^*(cA) \leq \sum \lambda(cI_n) = \sum c\lambda(I_n) = c \sum \lambda(I_n)$$

(the last equality is immediate from the definition of such sums as suprema of finite subsums; cf. 1.15.5 or §1.15, Exercise 4); then $(1/c)\lambda^*(cA) \leq \sum \lambda(I_n)$ (even if—especially if—the sum on the right side is $+\infty$), and varying the covering yields $(1/c)\lambda^*(cA) \leq \lambda^*(A)$. Thus $\lambda^*(cA) \leq c\lambda^*(A)$ and the inequality proves its own reverse: $\lambda^*(A) = \lambda^*((1/c)(cA)) \leq (1/c)\lambda^*(cA)$.

case 3: $c = -1$.

Then $cA = -A = \{-x : x \in A\}$ and $x \mapsto -x$ is an order-reversing bijection $\mathbb{R} \rightarrow \mathbb{R}$ such that $\lambda(-I) = \lambda(I)$ for all open intervals $I = (a, b)$; it follows easily that $\lambda^*(-A) = \lambda^*(A)$.

case 4: $c < 0$.

Then $cA = -[(-c)A]$, where $-c > 0$; by cases 3 and 2, $\lambda^*(cA) = \lambda^*((-c)A) = (-c)\lambda^*(A) = |c| \cdot \lambda^*(A)$. \diamond

The pattern of the foregoing proofs: a property of λ^* for arbitrary sets $A \subset \mathbb{R}$ is inferred from the analogous property of λ for open intervals I .

Here is a formal property of λ that does not carry over to λ^* : if I and J are bounded intervals such that $I \cap J = \emptyset$ and $I \cup J$ is an interval, then

$$\lambda(I \cup J) = \lambda(I) + \lambda(J);$$

for, if $I \cup J$ has endpoints $a \leq b$, then I and J are obtained by splitting $I \cup J$ at one of its points c , and (supposing I to be to the left of J) the asserted equality reduces to $b - a = (c - a) + (b - c)$. However, the implication

$$(*) \quad A \cap B = \emptyset \Rightarrow \lambda^*(A \cup B) = \lambda^*(A) + \lambda^*(B)$$

is in general false; a counterexample is given in §2.5, Exercise 2 (the Axiom of Choice is required!). This is expressed by saying that λ^* is not in general 'additive'.

The remedy is to restrict λ^* to a smaller class of sets on which λ^* is well-behaved; this is the subject of the next section.

Exercises

1. The union of a family of negligible sets need not be negligible; thus countability is essential in 2.1.7.

2. If $N \subset \mathbb{R}$ is negligible, then $\lambda^*(A \cup N) = \lambda^*(A)$ for every $A \subset \mathbb{R}$.

3. If A and B are subsets of \mathbb{R} such that $x < y$ for all $x \in A$ and $y \in B$, then $\lambda^*(A \cup B) = \lambda^*(A) + \lambda^*(B)$.

2.2. Measurable Sets

If a set $A \subset \mathbb{R}$ is expressed as a union $A = B \cup C$ of disjoint sets B and C , then the equation $\lambda^*(A) = \lambda^*(B) + \lambda^*(C)$ may fail (§2.5). However, if B and C are obtained by splitting A at some point $c \in \mathbb{R}$, then all is well:

2.2.1. Theorem. *If $c \in \mathbb{R}$ then c splits every subset A of \mathbb{R} additively, in the following sense: if*

$$B = A \cap (-\infty, c) \text{ and } C = A \cap [c, +\infty)$$

then $\lambda^(A) = \lambda^*(B) + \lambda^*(C)$.*

Proof. Suppose first that A is an open interval, say $A = (a, b)$. If $a < c < b$ then $B = (a, c)$, $C = [c, b)$ and the asserted equality reduces to $b - a = (c - a) + (b - c)$; if $c \leq a$ then $B = \emptyset$, $C = (a, b)$ and the equality is trivial, and similarly if $c \geq b$.

Now suppose $A \subset \mathbb{R}$ is arbitrary. Write $E = (-\infty, c)$, $E' = [c, +\infty)$, so that

$$B = A \cap E, \quad C = A \cap E'.$$

If (I_n) is a sequence of open intervals with $A \subset \bigcup I_n$, then $A \cap E \subset \bigcup (I_n \cap E)$ and the $I_n \cap E$ are bounded intervals (possibly empty), therefore

$$\lambda^*(A \cap E) \leq \sum \lambda(I_n \cap E);$$

similarly

$$\lambda^*(A \cap E') \leq \sum \lambda(I_n \cap E'),$$

and addition of these inequalities yields

$$\lambda^*(A \cap E) + \lambda^*(A \cap E') \leq \sum \lambda(I_n),$$

by the first paragraph of the proof. Varying the covering (I_n) ,

$$\lambda^*(A \cap E) + \lambda^*(A \cap E') \leq \lambda^*(A);$$

the reverse inequality follows from the subadditivity of λ^* (2.1.6). \diamond

This prompts a definition:

2.2.2. Definition. (C. Carathéodory). A set $E \subset \mathbb{R}$ is said to be **Lebesgue-measurable** (briefly, measurable) if E splits every subset of \mathbb{R} additively, in the sense that

$$\lambda^*(A) = \lambda^*(A \cap E) + \lambda^*(A \cap E') \quad \text{for all } A \subset \mathbb{R},$$

where $E' = \mathbb{R} - E$ is the complement of E in \mathbb{R} .

2.2.3. Remarks. (i) Since $A = (A \cap E) \cup (A \cap E')$, the inequality $\lambda^*(A) \leq \lambda^*(A \cap E) + \lambda^*(A \cap E')$ always holds by the subadditivity of λ^* ; to show that E is measurable, it suffices to verify the reverse inequality for all $A \subset \mathbb{R}$.

(ii) An example of a nonmeasurable set is given in §2.5.

(iii) For every real number c , the interval $E = (-\infty, c)$ is measurable (2.2.1).

(iv) E measurable $\Rightarrow E'$ measurable.

(v) Every negligible set is measurable. For, if $\lambda^*(E) = 0$ then, for every $A \subset \mathbb{R}$,

$$\lambda^*(A \cap E) + \lambda^*(A \cap E') = 0 + \lambda^*(A \cap E') \leq \lambda^*(A).$$

(vi) If E is measurable, then so is $E + c$ for every $c \in \mathbb{R}$.

{Proof: The function $f: \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = x + c$ is a bijection that preserves outer measure (2.1.10); thus, for every $A \subset \mathbb{R}$,

$$\begin{aligned} \lambda^*(A) &= \lambda^*(f^{-1}(A)) = \lambda^*(f^{-1}(A) \cap E) + \lambda^*(f^{-1}(A) \cap E') \\ &= \lambda^*[f(f^{-1}(A) \cap E)] + \lambda^*[f(f^{-1}(A) \cap E')] \\ &= \lambda^*(A \cap f(E)) + \lambda^*(A \cap f(E)'), \end{aligned}$$

therefore $f(E)$ is measurable.}

(vii) If E is measurable, then so is cE for every $c \in \mathbb{R}$.

{Proof: If $c = 0$ then $cE = \{0\}$ or \emptyset , so cE is measurable by (v). Assuming $c \neq 0$, let $f: \mathbb{R} \rightarrow \mathbb{R}$ be the bijection $f(x) = cx$. The measurability of $f(E)$ follows from 2.1.11 by an argument similar to the one for (vi).}

Summarizing, the class of measurable sets $E \subset \mathbb{R}$ includes the intervals $(-\infty, c)$ and the negligible sets, and is closed under complementation, translation and scalar multiplication. In the same vein:

2.2.4. Lemma. If E and F are measurable sets, then so are $E \cup F$, $E \cap F$ and $E - F$; if, moreover, $E \cap F = \emptyset$, then

$$\lambda^*(A \cap (E \cup F)) = \lambda^*(A \cap E) + \lambda^*(A \cap F) \quad \text{for all } A \subset \mathbb{R}.$$

Proof. It suffices to deal with $E \cup F$; the formulas $E \cap F = (E' \cup F')'$ and $E - F = E \cap F'$ then finish the job.

Let $A \subset \mathbb{R}$. We are to show that

$$\lambda^*(A) = \lambda^*(A \cap (E \cup F)) + \lambda^*(A \cap (E \cup F)').$$

Since E splits $A \cap (E \cup F)$ additively and

$$A \cap (E \cup F) \cap E = A \cap E, \quad A \cap (E \cup F) \cap E' = A \cap F \cap E',$$

we have

$$(*) \quad \lambda^*(A \cap (E \cup F)) = \lambda^*(A \cap E) + \lambda^*(A \cap F \cap E').$$

If E and F are disjoint (that is, $F \subset E'$), then $(*)$ yields the equation in the statement of the theorem. In general,

$$\begin{aligned} \lambda^*(A) &= \lambda^*(A \cap E) + \lambda^*(A \cap E') \\ &= \lambda^*(A \cap E) + \lambda^*(A \cap E' \cap F) + \lambda^*(A \cap E' \cap F') \\ &= \lambda^*(A \cap (E \cup F)) + \lambda^*(A \cap (E \cup F)') \end{aligned}$$

(the first equality because E splits A additively, the second because F splits $A \cap E'$ additively, the third by $(*)$). \diamond

2.2.5. Theorem. *If (E_n) is a sequence of measurable sets, then the sets*

$$E = \bigcup_{n=1}^{\infty} E_n, \quad F = \bigcap_{n=1}^{\infty} E_n$$

are also measurable. If, moreover, the E_n are pairwise disjoint, then

$$\lambda^*(A \cap E) = \sum_{n=1}^{\infty} \lambda^*(A \cap E_n)$$

for all $A \subset \mathbb{R}$.

Proof. Since $F = (\bigcup E'_n)'$, we need only consider E . The sets

$$E_1, E_1 \cup E_2, E_1 \cup E_2 \cup E_3, \dots$$

are measurable by the lemma, with union E ; thus, in proving E measurable, we can suppose that $E_1 \subset E_2 \subset E_3 \subset \dots$. Then the sets

$$E_1, E_2 - E_1, E_3 - E_2, \dots$$

are measurable, with union E ; changing notations again, we can suppose that the E_n are *pairwise disjoint*. Let $A \subset \mathbb{R}$. For each n ,

$$(i) \quad \lambda^* \left(A \cap \bigcup_{k=1}^n E_k \right) = \sum_{k=1}^n \lambda^*(A \cap E_k)$$

(by induction on the lemma). Also, $\bigcup_{k=1}^n E_k \subset E$, so $(\bigcup_{k=1}^n E_k)' \supset E'$, therefore

$$(ii) \quad \lambda^* \left(A \cap \left(\bigcup_{k=1}^n E_k \right)' \right) \geq \lambda^*(A \cap E');$$

adding (i), (ii) and citing the measurability of $\bigcup_{k=1}^n E_k$, we have

$$\lambda^*(A) \geq \sum_{k=1}^n \lambda^*(A \cap E_k) + \lambda^*(A \cap E').$$

Since n is arbitrary,

$$\begin{aligned} \lambda^*(A) &\geq \sum_{k=1}^{\infty} \lambda^*(A \cap E_k) + \lambda^*(A \cap E') \\ &\geq \lambda^*(A \cap E) + \lambda^*(A \cap E') \geq \lambda^*(A) \end{aligned}$$

(the second and third inequalities by the countable subadditivity of λ^*), whence equality throughout:

$$\lambda^*(A) = \lambda^*(A \cap E) + \lambda^*(A \cap E') = \sum_{k=1}^{\infty} \lambda^*(A \cap E_k) + \lambda^*(A \cap E').$$

The first equality shows that E is measurable; replacing A by $A \cap E$, the second inequality yields

$$\lambda^*(A \cap E) + 0 = \sum_{k=1}^{\infty} \lambda^*(A \cap E_k) + 0. \quad \diamond$$

2.2.6. Remark. For future use, we note that if X is any set, the theorem and its lemma are valid with λ^* replaced by any set function ρ on $\mathcal{P}(X)$ satisfying the conditions (1)–(4) of 2.1.6, that is,

- (1) $0 \leq \rho(A) \leq +\infty$ for all $A \subset X$,
- (2) $\rho(\emptyset) = 0$,
- (3) $A \subset B \Rightarrow \rho(A) \leq \rho(B)$,
- (4) $\rho(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \rho(A_n)$ for every sequence (A_n) of subsets of X .

2.2.7. Definition. A function ρ satisfying the above conditions (1)–(4) is called a (Carathéodory) **outer measure** on X . A set $E \subset X$ is then called **ρ -negligible** if $\rho(E) = 0$, and **ρ -measurable** if it splits every set $A \subset X$ additively (in the sense of 2.2.2).

{CAUTION: It can happen that \emptyset and X are the only ρ -measurable sets (Exercise 1).}

We now pursue the consequences of Theorem 2.2.5:

2.2.8. Corollary. *Every interval in \mathbb{R} is measurable.*

Proof. Every interval I is the union of a sequence of bounded intervals; by the preceding theorem, we can suppose that I is bounded. If $I = [a, b)$ then $I = (-\infty, b) - (-\infty, a)$ is the difference of two measurable sets (2.2.3), hence is measurable (2.2.4). Since singletons are negligible, hence measurable, it then follows that $[a, b] = [a, b) \cup \{b\}$, $(a, b] = [a, b] - \{a\}$ and $(a, b) = [a, b] - \{a\}$ are also measurable. \diamond

2.2.9. *Definition.* We write $\mathcal{M} = \mathcal{M}(\lambda^*)$ for the set of all Lebesgue-measurable subsets of \mathbb{R} . Since $\lambda^*(I) = \lambda(I)$ for all bounded intervals I (2.1.6), we may consistently define a function

$$\lambda : \mathcal{M} \rightarrow [0, +\infty]$$

by the formula $\lambda(E) = \lambda^*(E)$ for all $E \in \mathcal{M}$, that is, $\lambda = \lambda^*|_{\mathcal{M}}$ (the restriction of λ^* to \mathcal{M}); this set function λ is called **Lebesgue measure** on \mathbb{R} .

The remaining corollaries depend only on Theorem 2.2.5 and its lemma, so they are valid for an outer measure ρ and its restriction to the class of ρ -measurable sets (2.2.6, 2.2.7). Lebesgue measure is ‘countably additive’ in the following sense:

2.2.10. *Corollary.* If (E_n) is a sequence of pairwise disjoint measurable sets, then

$$\lambda\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \lambda(E_n).$$

Proof. In 2.2.5, put $A = E$ (or $A = \mathbb{R}$). \diamond

2.2.11. *Corollary.* If (F_n) is an increasing sequence of measurable sets with union F (briefly $F_n \uparrow F$), then $\lambda^*(A \cap F_n) \uparrow \lambda^*(A \cap F)$ for all $A \subset \mathbb{R}$; in particular, $\lambda(F_n) \uparrow \lambda(F)$.

Proof. Let $E_1 = F_1$ and $E_n = F_n - F_{n-1}$ for $n > 1$. The E_n are pairwise disjoint measurable sets with union F , therefore (2.2.5)

$$\begin{aligned} \lambda^*(A \cap F) &= \sum_{n=1}^{\infty} \lambda^*(A \cap E_n) = \sup_n \left(\sum_{k=1}^n \lambda^*(A \cap E_k) \right) \\ &= \sup_n \lambda^* \left(A \cap \bigcup_{k=1}^n E_k \right) = \sup_n \lambda^*(A \cap F_n) \end{aligned}$$

for all $A \subset \mathbb{R}$. \diamond

2.2.12. *Corollary.* If (G_n) is a decreasing sequence of measurable sets with intersection G (briefly $G_n \downarrow G$) and if $\lambda(G_1) < +\infty$, then $\lambda^*(A \cap G_n) \downarrow \lambda^*(A \cap G)$ for all $A \subset \mathbb{R}$; in particular, $\lambda(G_n) \downarrow \lambda(G)$.

Proof. Writing $F_n = G_1 - G_n$ and $F = G_1 - G$, we have $F_n \uparrow F$. Let $A \subset \mathbb{R}$. By the preceding corollary, $\lambda^*(A \cap F_n) \uparrow \lambda^*(A \cap F)$, where $\lambda^*(A \cap F) \leq \lambda^*(G_1) < +\infty$. For all n ,

$$G_1 = F_n \cup G_n = F \cup G \quad \text{and} \quad F_n \cap G_n = F \cap G = \emptyset,$$

therefore (2.2.4)

$$\begin{aligned}\lambda^*(A \cap G_1) &= \lambda^*(A \cap F_n) + \lambda^*(A \cap G_n), \\ \lambda^*(A \cap G_1) &= \lambda^*(A \cap F) + \lambda^*(A \cap G).\end{aligned}$$

By finiteness,

$$\begin{aligned}\lambda^*(A \cap G_n) &= \lambda^*(A \cap G_1) - \lambda^*(A \cap F_n), \\ \lambda^*(A \cap G) &= \lambda^*(A \cap G_1) - \lambda^*(A \cap F),\end{aligned}$$

whence $\lambda^*(A \cap G_n) \downarrow \lambda^*(A \cap G)$. \diamond

Exercises

1. Give an example of an outer measure ρ on a set X , for which \emptyset and X are the only ρ -measurable sets.

{Hint: Consider a two-point set $X = \{a, b\}$.}

2. Every open set and every closed set in \mathbb{R} is (Lebesgue) measurable.

{Hint: 2.2.8 and 2.2.5.}

3. Let (E_n) be a sequence of measurable sets in \mathbb{R} .

(i) The sets

$$\begin{aligned}F &= \liminf E_n = \{x : x \in E_n \text{ ultimately}\} \\ E &= \limsup E_n = \{x : x \in E_n \text{ frequently}\}\end{aligned}$$

are also measurable.

{Hint: For the terminology, see 1.16.4. If $F_n = \bigcap_{k \geq n} E_k$, then $F_n \uparrow F$.}

(ii) $\lambda(\liminf E_n) \leq \liminf \lambda(E_n)$.

(iii) If $\lambda(\bigcup E_n) < +\infty$, then $\lambda(\limsup E_n) \geq \limsup \lambda(E_n)$.

(iv) The inequality $\lambda(\limsup E_n) \geq \limsup \lambda(E_n)$ is in general false.

{Hint: Try $E_n = [n, n+1]$.}

4. Let (A_n) be any sequence of subsets of \mathbb{R} .

(i) If $A_n \uparrow A$ then $\lambda^*(A_n) \uparrow \lambda^*(A)$.

(ii) $\lambda^*(\liminf A_n) \leq \liminf \lambda^*(A_n)$.

{Hint: (i) For every subset S of \mathbb{R} , there exists a measurable set E such that $S \subset E$ and every measurable subset of $E - S$ is negligible.¹}

5. For $A \subset [a, b] \subset \mathbb{R}$, $\lambda^*(A)$ is also called the *exterior measure* of A in $[a, b]$ and is denoted $\lambda_e(A)$; the *interior measure* of A in $[a, b]$, denoted $\lambda_i(A)$, is defined by the formula

$$\lambda_i(A) = (b - a) - \lambda_e([a, b] - A).$$

¹ Cf. the author, *Measure and integration* [Chelsea, New York, 1970], §8, Theorem 1.

In general, $\lambda_i(A) \leq \lambda_e(A)$; the set A is measurable if and only if $\lambda_i(A) = \lambda_e(A)$.

6. For a sequence of functions $f_n : X \rightarrow \overline{\mathbb{R}}$ the functions $\limsup f_n$ and $\liminf f_n$ are defined by the formulas

$$\begin{aligned}(\limsup f_n)(x) &= \limsup f_n(x), \\ (\liminf f_n)(x) &= \liminf f_n(x).\end{aligned}$$

If (E_n) is a sequence of subsets of a set X and if φ denotes characteristic function, then

$$(i) \quad \limsup \varphi_{E_n} = \varphi_{\limsup E_n},$$

$$(ii) \quad \liminf \varphi_{E_n} = \varphi_{\liminf E_n}.$$

(For the notations on the right side, see Exercise 3.)

2.3. Cantor Set: An Uncountable Set of Measure Zero

The Cantor set Γ is a negligible, closed subset of $[0, 1]$ with cardinality c ($= \text{card } \mathbb{R}$); it is constructed by deleting the open middle third $(\frac{1}{3}, \frac{2}{3})$ of $[0, 1]$, then deleting the open middle thirds of the two closed intervals that remain, 'and so on' (a set-theoretic passage to the limit). Establishing the right notation is 99% of the battle.

If $A = [a, b]$ is a nondegenerate closed interval, we write $r(A)$ ('the rest of A ') for what is left of A after deleting the open middle third:

$$r(A) = [a, a + \frac{1}{3}(b - a)] \cup [b - \frac{1}{3}(b - a), b],$$

a set whose Lebesgue measure is $\frac{2}{3}\lambda(A)$. More generally, if $A = A_1 \cup \dots \cup A_k$ is a finite union of pairwise disjoint, nondegenerate closed intervals A_i , we define

$$r(A) = r(A_1) \cup \dots \cup r(A_k),$$

where the sets $r(A_i)$ have the meaning defined above; the right hand side is a set of the same sort as A , so the operation r on such sets A can be iterated, $r^n(A)$ ($n = 1, 2, 3, \dots$) being defined recursively by the formulas

$$r^1(A) = r(A), \quad r^{n+1}(A) = r(r^n(A)).$$

The following properties of this operation are easily checked:

- (1) $A \supset r(A) \supset r^2(A) \supset \dots$
- (2) $\lambda(r^n(A)) = (\frac{2}{3})^n \lambda(A)$, where λ is Lebesgue measure.
- (3) $r^n(A)$ is a closed set¹ (it is the union of finitely many closed intervals).

¹ First course, §4.2.

We write $r^\infty(A) = \bigcap_{n=1}^\infty r^n(A)$ for the intersection of the decreasing sequence $(r^n(A))$.

(4) $r^\infty(A)$ is a closed set, and $\lambda(r^\infty(A)) \leq \lambda(r^n(A))$ shows that its Lebesgue measure is 0.

The Cantor set is the result of applying this machinery to the closed unit interval:

2.3.1. *Definition.* The set $\Gamma = r^\infty([0, 1])$ is called the **Cantor set**; it is a closed set in \mathbb{R} with $\lambda(\Gamma) = 0$.

Let $I = [0, 1]$. It is useful to refine the foregoing notations. We have

$$r(I) = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1] = I_0 \cup I_1,$$

where I_0 is the ‘left third’ of I , and I_1 is the ‘right third’ of I . In turn,

$$r^2(I) = I_{00} \cup I_{01} \cup I_{10} \cup I_{11},$$

where, for example, I_{10} is the left third of I_1 . For every n -ple $\alpha = (\alpha_1, \dots, \alpha_n)$ with $\alpha_1, \dots, \alpha_n \in \{0, 1\}$, we recursively define

$$I_\alpha = I_{\alpha_1 \alpha_2 \dots \alpha_n}$$

to be the left third of $I_{\alpha_1 \alpha_2 \dots \alpha_{n-1}}$ if $\alpha_n = 0$, and the right third if $\alpha_n = 1$.

Let us call $\alpha = (\alpha_1, \dots, \alpha_n)$ an *index of rank n* and write $|\alpha| = n$. For indices α, β let us write $\alpha \leq \beta$ in case $|\alpha| \leq |\beta|$ and $\alpha_i = \beta_i$ for $i = 1, 2, \dots, |\alpha|$ (so to speak, α is the ‘initial $|\alpha|$ -ple’ of β). The following properties are easily verified:

(5) If $|\alpha| = n$ then I_α is one of the 2^n closed intervals that make up $r^n(I)$, and $\lambda(I_\alpha) = (\frac{1}{3})^n$.

(6) If $\alpha \leq \beta$ then $I_\alpha \supset I_\beta$.

(7) If $\alpha^1, \alpha^2, \alpha^3, \dots$ is a sequence of indices with $\alpha^1 \leq \alpha^2 \leq \alpha^3 \leq \dots$ and if $|\alpha^n| = n$ for all n , then

$$\bigcap_{n=1}^\infty I_{\alpha^n} = \{y\}$$

for some $y \in \Gamma$. {Sketch of proof: The intersection is a singleton $\{y\}$ by the theorem on nested intervals (1.8.27) and $I_{\alpha^n} \subset r^n(I)$ for all n , therefore $y \in r^\infty(I) = \Gamma$.}

(8) If $|\alpha| = |\beta|$ and $\alpha \neq \beta$ then $I_\alpha \cap I_\beta = \emptyset$. {Consider the first coordinate in which α and β differ; clearly I_α and I_β are contained in disjoint ‘thirds’ of some interval.}

It remains to show that Γ has cardinality $c = 2^{\aleph_0}$. The set

$$M = \prod_{n=1}^\infty \{0, 1\}$$

(the product of \aleph_0 copies of the 2-element set $\{0, 1\}$) has cardinality c (1.13.15), so it will suffice to exhibit a bijection

$$f : M \rightarrow \Gamma.$$

Let $x \in M$, say $x = (x_1, x_2, x_3, \dots)$. For every positive integer n , let $\alpha^n(x)$ be the initial n 'ple of x , that is,

$$\alpha^n(x) = (x_1, \dots, x_n).$$

In the notations preceding (5),

$$\alpha^1(x) \leq \alpha^2(x) \leq \alpha^3(x) \leq \dots$$

and $|\alpha^n(x)| = n$. By (7), there is a unique point y of Γ such that

$$\bigcap_{n=1}^{\infty} I_{\alpha^n(x)} = \{y\}.$$

We define $f(x) = y$. Thus, $f : M \rightarrow \Gamma$ is defined by the condition

$$\bigcap_{n=1}^{\infty} I_{\alpha^n(x)} = \{f(x)\}.$$

2.3.2. Theorem. *Let Γ be the Cantor set (2.3.1), M the product of \aleph_0 copies of $\{0, 1\}$, and $f : M \rightarrow \Gamma$ the mapping defined above. Then f is bijective. In particular, $\text{card } \Gamma = c$.*

Proof. Injectivity. If $x \neq x'$ then $\alpha^n(x) \neq \alpha^n(x')$ for some n , and $f(x) \neq f(x')$ follows at once from (8).

Surjectivity. Given $y \in \Gamma$, we construct a point $x = (x_1, x_2, x_3, \dots)$ in M such that $f(x) = y$. The coordinates x_n are defined as follows. We know that $y \in r^n(I)$ for all n . In particular, $y \in r(I) = I_0 \cup I_1$; let x_1 be the element of $\{0, 1\}$ such that $y \in I_{x_1}$. Since $y \in r^2(I)$, either $y \in I_{x_1 0}$ or $y \in I_{x_1 1}$; let x_2 be the element of $\{0, 1\}$ such that $y \in I_{x_1 x_2}$, and so on. \diamond

Exercises

1. The Cantor set contains no interior point.

2. Every $y \in \Gamma$ is a limit point of Γ , that is, there exists a sequence $y_n \in \Gamma$ such that $y_n \neq y$ and $y_n \rightarrow y$.

{Hint: As in the proof of 2.3.2, let x be the point of M such that $y \in I_{\alpha^n(x)}$ for all n . For each n , let y_n be an endpoint of the interval $I_{\alpha^n(x)}$ such that $y \neq y_n$.}

2.4. Borel Sets, Regularity

The properties of the class of measurable sets described in §2.2 prompt the following definition:

2.4.1. *Definition.* Let X be a set. A set \mathcal{S} of subsets of X is called a σ -**algebra** (on X) if (1) $\emptyset \in \mathcal{S}$, (2) $E \in \mathcal{S} \Rightarrow E' \in \mathcal{S}$, and (3) $E_n \in \mathcal{S}$ ($n = 1, 2, 3, \dots$) $\Rightarrow \bigcup E_n \in \mathcal{S}$. In other words, \mathcal{S} contains the empty set and it is closed under complementation and denumerable unions.

If (3) is replaced by the weaker condition (3') $A, B \in \mathcal{S} \Rightarrow A \cup B \in \mathcal{S}$ (so that \mathcal{S} is closed under finite unions), then \mathcal{S} is called an **algebra** of subsets of X .

2.4.2. *Remarks.* If \mathcal{S} is an algebra of subsets of X , then:

- (i) $X \in \mathcal{S}$ (because $X = \emptyset'$);
- (ii) $E, F \in \mathcal{S} \Rightarrow E \cap F = (E' \cup F')' \in \mathcal{S}$;
- (iii) $E, F \in \mathcal{S} \Rightarrow E - F = E \cap F' \in \mathcal{S}$.
- (iv) Every σ -algebra is an algebra (consider $E_1 = A$ and $E_n = B$ for $n \geq 2$).
- (v) If \mathcal{S} is a σ -algebra and $E_n \in \mathcal{S}$ ($n = 1, 2, 3, \dots$), then $\bigcap E_n = (\bigcup E'_n)' \in \mathcal{S}$.
- (iv) If \mathcal{S} and \mathcal{T} are algebras (σ -algebras) on X then so is

$$\mathcal{S} \cap \mathcal{T} = \{A \subset X : A \in \mathcal{S} \text{ \& } A \in \mathcal{T}\};$$

more generally, the intersection of any family of algebras (σ -algebras) on X is an algebra (a σ -algebra).

2.4.3. *Examples.* (i) $\mathcal{P}(X)$ is the largest σ -algebra on X , $\{\emptyset, X\}$ the smallest.

(ii) The set \mathcal{M} of Lebesgue-measurable sets is a σ -algebra on \mathbb{R} (2.2.3–2.2.5). More generally,

(iii) if ρ is any outer measure on a set X , then the set of all ρ -measurable sets is a σ -algebra on X (2.2.6–2.2.7).

Every set of sets is contained in a minimal σ -algebra:

2.4.4. *Definition.* Let X be a set, $\mathcal{E} \subset \mathcal{P}(X)$ a set of subsets of X . There exist σ -algebras on X that contain \mathcal{E} (for example, $\mathcal{P}(X)$); the intersection of them all, denoted

$$\mathcal{S}(\mathcal{E}) = \bigcap \{\mathcal{S} : \mathcal{S} \text{ a } \sigma\text{-algebra on } X \text{ with } \mathcal{E} \subset \mathcal{S}\},$$

is also a σ -algebra (2.4.2, (vi)) containing \mathcal{E} , called the σ -algebra **generated** by \mathcal{E} , and it is characterized by the following three properties: (1) $\mathcal{S}(\mathcal{E})$ is a σ -algebra on X , (2) $\mathcal{E} \subset \mathcal{S}(\mathcal{E})$, and (3) if \mathcal{S} is a σ -algebra on X such that $\mathcal{E} \subset \mathcal{S}$, then $\mathcal{S}(\mathcal{E}) \subset \mathcal{S}$.

The algebra $\mathcal{A}(\mathcal{E})$ on X *generated* by \mathcal{E} is defined analogously, with “ σ -algebra” replaced by “algebra” in the preceding definition.

In many situations, the σ -algebra of Lebesgue-measurable sets (whose definition depends on the outer measure λ^*) can be effectively replaced by the following smaller σ -algebra (defined without reference to an outer measure):

2.4.5. *Definition.* Let \mathcal{I} be the set of all open intervals (a, b) in \mathbb{R} . The sets in the σ -algebra $\mathcal{S}(\mathcal{I})$ generated by \mathcal{I} are called the **Borel sets**¹ (or ‘Baire sets’²) of \mathbb{R} . This σ -algebra is also denoted $\mathcal{B}(\mathbb{R})$, briefly \mathcal{B} . (The use of *open* intervals here is not crucial; see Exercise 1.)

2.4.6. *Theorem.* Every Borel set in \mathbb{R} is Lebesgue-measurable.

Proof. The set \mathcal{M} of Lebesgue-measurable sets is a σ -algebra containing the set \mathcal{I} of all open intervals, therefore $\mathcal{M} \supset \mathcal{S}(\mathcal{I}) = \mathcal{B}$. \diamond

2.4.7. *Remark.* The inclusion $\mathcal{B} \subset \mathcal{M}$ of the preceding theorem is proper: there exist measurable sets that are not Borel (see Exercise 2).

The concept of Borel set can be extended to situations vastly more general than the real line by liberating it from its dependence on intervals. To this end, let us review³ some elementary ‘topology’ of \mathbb{R} :

2.4.8. *Definition.* Let $a \in A \subset \mathbb{R}$. We say that a is an *interior point* of A (or that A is a *neighborhood* of a) if there exists a real number $r > 0$ such that $(a - r, a + r) \subset A$ (in other words, $|x - a| < r \Rightarrow x \in A$). If every point of A is an interior point, then A is said to be *open*. Convention: \emptyset is an open set in \mathbb{R} .

2.4.9. *Remarks.* (i) Every open interval is an open set.

(ii) \emptyset and \mathbb{R} are open sets in \mathbb{R} .

(iii) If A and B are open sets in \mathbb{R} , then so is $A \cap B$.

(iv) The union of any family of open sets is open.

(v) A subset of \mathbb{R} is open if and only if it is the union of a family of open intervals (immediate from (i), (iv) and the definition of open set). Better yet:

2.4.10. *Lemma.* Every open set in \mathbb{R} is the union of a sequence of open intervals.

Proof. Let U be an open set in \mathbb{R} . If $x \in U$, there exists an open interval (a, b) with $x \in (a, b) \subset U$, and we can suppose that a and b are rational numbers. Let \mathcal{I}_r be the set of all open intervals with rational endpoints; by the preceding remark, U is the union of a subset of \mathcal{I}_r , so it will suffice to observe that \mathcal{I}_r is countable. Indeed, the mapping $\mathcal{I}_r \rightarrow \mathbb{Q} \times \mathbb{Q}$ that sends an open interval $(a; b) \in \mathcal{I}_r$ to the ordered pair of

¹ Émile Borel (1871-1956).

² René Baire (1874-1932).

³ *First course*, p. 65, §4.3.

rational numbers $(a, b) \in \mathbb{Q} \times \mathbb{Q}$ is injective, and $\mathbb{Q} \times \mathbb{Q}$ is denumerable (1.10.9–1.10.10). \diamond

2.4.11. Theorem. *Every open set in \mathbb{R} is a Borel set. In fact, \mathcal{B} is the σ -algebra generated by the set of all open sets.*

Proof. Let \mathcal{O} be the set of all open sets in \mathbb{R} . The first assertion, that $\mathcal{O} \subset \mathcal{B}$, is immediate from the lemma, and it follows that $\mathcal{S}(\mathcal{O}) \subset \mathcal{B}$; on the other hand, $\mathcal{I} \subset \mathcal{O}$, therefore $\mathcal{B} = \mathcal{S}(\mathcal{I}) \subset \mathcal{S}(\mathcal{O})$. \diamond

Recall that $\mathcal{B} \subset \mathcal{M}$ (2.4.6); even though \mathcal{M} is ‘much larger’ than \mathcal{B} , we shall see below that Lebesgue measure $\lambda = \lambda^*|_{\mathcal{M}}$ is characterized by its restriction to \mathcal{B} . Both λ and $\lambda|_{\mathcal{B}}$ are special cases of the following concept:

2.4.12. Definition. Let X be a set and \mathcal{S} a σ -algebra of subsets of X . A **measure** on \mathcal{S} is a function $\mu : \mathcal{S} \rightarrow [0, +\infty]$ such that (1) $\mu(\emptyset) = 0$, and (2) μ is countably additive in the sense that $\mu(\bigcup E_n) = \sum \mu(E_n)$ for every sequence (E_n) of pairwise disjoint sets in \mathcal{S} . The triple (X, \mathcal{S}, μ) is then called a **measure space**; we also say that ‘ μ is a measure on X ’.

2.4.13. Examples. (i) The triple $(\mathbb{R}, \mathcal{M}, \lambda)$ is a measure space, as is $(\mathbb{R}, \mathcal{B}, \lambda|_{\mathcal{B}})$.

(ii) If ρ is any outer measure on a set X and \mathcal{S} is the σ -algebra of all ρ -measurable sets (2.4.3, (iii)), then $(X, \mathcal{S}, \rho|_{\mathcal{S}})$ is a measure space by the remark preceding 2.2.10.

We know that $\mu = \lambda|_{\mathcal{B}}$ is a measure on \mathcal{B} such that $\lambda(I) = b - a$ for every open interval $I = (a, b)$. In a later chapter we shall see that any two measures on $\mathcal{B} = \mathcal{S}(\mathcal{I})$ that agree on the set \mathcal{I} of open intervals are identical (4.6.10), so that, in particular, μ is uniquely determined by the property $\mu((a, b)) = b - a$.

In this section, we characterize Lebesgue measure λ on \mathcal{M} in a similar way: a measure on \mathcal{M} that agrees with λ on Borel sets is identical with λ on \mathcal{M} (2.4.16 below). The key to this result is the fact that every Lebesgue-measurable set is ‘almost’ a Borel set (2.4.15 below), and the basic technique is the following approximation theorem:

2.4.14. Theorem. *For every set $A \subset \mathbb{R}$,*

$$\lambda^*(A) = \inf\{\lambda(U) : A \subset U, U \text{ open}\}.$$

Moreover, there exists a decreasing sequence (U_n) of open sets such that $A \subset U_n$ and $\lambda(U_n) \downarrow \lambda^(A)$.*

Proof. Write α for the infimum on the right side. For every open set U with $A \subset U$, we have $\lambda^*(A) \leq \lambda^*(U) = \lambda(U)$, therefore $\lambda^*(A) \leq \alpha$. On the other hand, if (I_n) is a sequence of open intervals such that $A \subset \bigcup I_n$,

then $U = \bigcup I_n$ is an open set containing A , therefore

$$\alpha \leq \lambda(U) \leq \sum \lambda(I_n);$$

varying the covering (I_n) of A , we have $\alpha \leq \lambda^*(A)$. This proves the first assertion of the theorem.

Choose a sequence (V_n) of open sets such that $A \subset V_n$ and $\lambda^*(A) = \inf \lambda(V_n)$. {If $\alpha = +\infty$, let $V_n = \mathbb{R}$ for all n ; if $\alpha < +\infty$ then, for each n , there exists an open set $V_n \supset A$ such that $\lambda(V_n) \leq \lambda^*(A) + 1/n$ for all n .} Let $U_n = V_1 \cap \dots \cap V_n$. Then (U_n) is a decreasing sequence of open sets such that $A \subset U_n \subset V_n$ for all n , and $\lambda^*(A) \leq \lambda(U_n) \leq \lambda(V_n)$ shows that $\lambda^*(A) = \inf \lambda(U_n)$, thus $\lambda(U_n) \downarrow \lambda^*(A)$. \diamond

2.4.15. Corollary. *If E is a Lebesgue-measurable set, then there exist Borel sets F and G such that*

$$F \subset E \subset G \text{ and } \lambda(G - F) = 0.$$

Proof. It suffices to find (i) a Borel set G such that $E \subset G$ and $\lambda(G - E) = 0$, and (ii) a Borel set F such that $F \subset E$ and $\lambda(E - F) = 0$; then $G - F = (G - E) \cup (E - F)$ shows that $\lambda(G - F) = 0$.

(i) Suppose first that $\lambda(E) < +\infty$. By the theorem, there exists a decreasing sequence (U_n) of open sets such that $E \subset U_n$ and $\lambda(U_n) \downarrow \lambda(E)$. Discarding a finite number of terms, we can suppose that $\lambda(U_1) < +\infty$. Then $G = \bigcap U_n$ is a Borel set containing E , $U_n \downarrow G$, and $\lambda(U_n) \downarrow \lambda(G)$ by 2.2.12, thus $\lambda(G) = \inf \lambda(U_n) = \lambda(E) < +\infty$. From $G = (G - E) \cup E$ we have

$$\lambda(G) = \lambda(G - E) + \lambda(E),$$

therefore (by finiteness) $\lambda(G - E) = \lambda(G) - \lambda(E) = 0$.

In the general case, $E = \bigcup E_n$ with (E_n) a sequence of measurable sets such that $\lambda(E_n) < +\infty$ for all n (for example, let $E_n = E \cap (-n, n)$). Applying the preceding argument to each E_n , we have a sequence (G_n) of Borel sets such that $E_n \subset G_n$ and $\lambda(G_n - E_n) = 0$ for all n . Then $G = \bigcup G_n$ is a Borel set containing E , and

$$G - E = \bigcup G_n - \bigcup E_n \subset \bigcup (G_n - E_n);$$

since $\lambda(G_n - E_n) = 0$ for all n , it follows that $\lambda(G - E) = 0$.

(ii) Applying (i) to the Lebesgue-measurable set $E' = \mathbb{R} - E$, there exists a Borel set B such that $E' \subset B$ and $\lambda(B - E') = 0$, that is, $\lambda(B \cap E) = 0$; then $F = B'$ is a Borel set such that $F \subset E$ and $\lambda(E - F) = \lambda(E \cap B) = 0$. \diamond

2.4.16. Corollary. *If ν is a measure on \mathcal{M} such that $\nu(B) = \lambda(B)$ for every Borel set B , then $\nu = \lambda$ on \mathcal{M} .⁴*

⁴In fact, it suffices to assume that $\nu(I) = \lambda(I)$ for every open interval I (4.6.10).

Proof. Let $E \in \mathcal{M}$. With notations as in the preceding corollary, we have $\nu(F) = \lambda(F)$, $\nu(G) = \lambda(G)$ and $\nu(G - F) = \lambda(G - F) = 0$. From $F \subset E \subset G = F \cup (G - F)$, we have

$$\nu(F) \leq \nu(E) \leq \nu(G) = \nu(F) + \nu(G - F) = \nu(F) + 0,$$

thus $\nu(F) = \nu(E) = \nu(G)$, and similarly $\lambda(F) = \lambda(E) = \lambda(G)$; in particular, $\nu(E) = \nu(F) = \lambda(F) = \lambda(E)$. \diamond

2.4.17. *Remark.* With notations as in the preceding corollary, let $N = E - F$; then $E = F \cup N$ with F Borel and $N \subset G - F$ negligible. In this sense, every Lebesgue-measurable set E is ‘almost Borel’.

The property of Lebesgue outer measure proved in 2.4.14,

$$\lambda^*(A) = \inf\{\lambda(U) : A \subset U, U \text{ open}\},$$

is called **outer regularity**; so to speak, every set $A \subset \mathbb{R}$ can be ‘approximated from the outside’ (more appropriately, ‘from above’) by open sets. The following property of Lebesgue measure, called **inner regularity**, says that every measurable set can be ‘approximated from the inside’ by compact sets⁵:

2.4.18. *Theorem.* If E is a Lebesgue-measurable set, then

$$\lambda(E) = \sup\{\lambda(K) : K \subset E, K \text{ compact}\}.$$

Proof. Let α be the supremum on the right side. The inequality $\lambda(E) \geq \alpha$ is immediate from the monotonicity of λ .

Suppose first that E is bounded, say $E \subset C = [a, b]$. Let $\epsilon > 0$. Since $\lambda(C - E) \leq \lambda(C) < +\infty$, we know from outer regularity that there exists an open set U such that $C - E \subset U$ and

$$\lambda(U) < \lambda(C - E) + \epsilon = \lambda(C) - \lambda(E) + \epsilon.$$

The set $K = C - U = C \cap U'$ is closed and bounded, that is, compact, and from $C - E \subset U$ we see that $K = C - U \subset E$, therefore $\lambda(K) \leq \alpha$ by the definition of α . Moreover,

$$C \subset U \cup (C - U) = U \cup K,$$

therefore

$$\lambda(C) \leq \lambda(U) + \lambda(K) < [\lambda(C) - \lambda(E) + \epsilon] + \lambda(K),$$

thus $\lambda(E) < \lambda(K) + \epsilon \leq \alpha + \epsilon$; since ϵ is arbitrary, $\lambda(E) \leq \alpha$, which completes the proof that $\lambda(E) = \alpha$ in case E is bounded.

Now let E be an arbitrary measurable set. If $\alpha = +\infty$ then $\lambda(E) = \alpha$ is forced by the inequality $\lambda(E) \geq \alpha$. Suppose $\alpha < +\infty$. Let $E_n =$

⁵ A subset of \mathbb{R} is said to be compact if it is closed and bounded (*First course*, p. 77, 4.5.6; cf. 6.1.9 below).

$E \cap (-n, n)$ ($n = 1, 2, 3, \dots$); the sets E_n are measurable and $E_n \uparrow E$, therefore $\lambda(E_n) \uparrow \lambda(E)$ (2.2.11). Moreover, each E_n is bounded, so by the preceding paragraph there exists a compact set $K_n \subset E_n$ such that $\lambda(E_n) < \lambda(K_n) + 1/n$. Since $K_n \subset E_n \subset E$, we have $\lambda(K_n) \leq \alpha$, therefore $\lambda(E_n) < \alpha + 1/n$ for all n , and passage to the limit yields $\lambda(E) \leq \alpha$. \diamond

The existence of sets $A \subset \mathbb{R}$ such that $\lambda^*(A) > \sup\{\lambda(K) : K \subset A, K \text{ compact}\}$ is discussed in §2.5, Exercise 1.

Exercises

1. If \mathcal{E} is the set of all closed (left-closed and right-open, left-open and right-closed) intervals in \mathbb{R} , then $\mathcal{S}(\mathcal{E}) = \mathcal{B}$.

2. There are c Borel sets in \mathbb{R} , and 2^c Lebesgue-measurable sets.

{Sketch: Since $\mathcal{M} \subset \mathcal{P}(\mathbb{R})$ and $\text{card } \mathcal{P}(\mathbb{R}) = 2^c$ (1.13.13), we have $\text{card } \mathcal{M} \leq 2^c$; on the other hand, the Cantor set has cardinality c (2.3.2) and its subsets are all measurable (2.2.3, (v)), consequently $\text{card } \mathcal{M} \geq 2^c$, and $\text{card } \mathcal{M} = 2^c$ by the Schröder-Bernstein theorem (1.12.6). There are at least c Borel sets, for example the singletons $\{a\}$, $a \in \mathbb{R}$, thus $\text{card } \mathcal{B} \geq c$; the proof that $\text{card } \mathcal{B} \leq c$ is more subtle (cf. Exercise 7).}

3. Let X be any set.

(i) For every subset $A \subset X$, define $\delta(A) = +\infty$ if A is infinite, and $\delta(A) = n$ if A is finite and has n elements. Then $(X, \mathcal{P}(X), \delta)$ is a measure space and δ is the unique measure on $\mathcal{P}(X)$ such that $\delta(\{x\}) = 1$ for all $x \in X$. (This is in stark contrast to Lebesgue measure, for which every singleton is negligible.) Such measures δ are called *discrete*.

(ii) Define $\tau(\emptyset) = 0$ and $\tau(A) = +\infty$ for every nonempty set $A \subset X$. Then $(X, \mathcal{P}(X), \tau)$ is a measure space (called, justifiably, *trivial*).

(iii) Fix a point $x \in X$ and, for $A \subset X$, define $\varepsilon_x(A) = 1$ if $x \in A$ and $\varepsilon_x(A) = 0$ if $x \notin A$. Then $(X, \mathcal{P}(X), \varepsilon_x)$ is a measure space; ε_x is called the *unit point mass* (or Dirac measure⁶) concentrated at x . (In a subtle way, all measures are made up of Dirac measures; the trick is in putting them together.⁷)

4. Let \mathcal{E} be a set of subsets of \mathbb{R} , that is, $\mathcal{E} \subset \mathcal{P}(\mathbb{R})$.

(i) One writes \mathcal{E}_δ for the set of all $S \subset \mathbb{R}$ such that $S = \bigcap E_n$ for some sequence (E_n) of sets in \mathcal{E} . If \mathcal{O} is the set of all open sets in \mathbb{R} , a set $E \in \mathcal{O}_\delta$ is called a G_δ (G for 'Gebiet', δ for 'Durchschnitt'). Every G_δ is a Borel set. For any \mathcal{E} , $(\mathcal{E}_\delta)_\delta = \mathcal{E}_\delta$. If \mathcal{C} is the set of all closed sets in \mathbb{R} , then $\mathcal{C}_\delta = \mathcal{C}$.

⁶ Paul Dirac.

⁷ N. Bourbaki, *Intégration*, Ch. 5 [2nd. edn., Hermann, Paris, 1967], §3, No. 1, p. 17, Example 2.

(ii) One writes \mathcal{E}_σ for the set of all $S \subset \mathbb{R}$ such that $S = \bigcup E_n$ for some sequence (E_n) of sets in \mathcal{E} . If \mathcal{C} is the set of all closed sets, a set $E \in \mathcal{C}_\sigma$ is called an F_σ (F for 'fermé', σ for 'somme'). Every F_σ is a Borel set. For any \mathcal{E} , $(\mathcal{E}_\sigma)_\sigma = \mathcal{E}_\sigma$. If \mathcal{O} is the set of all open sets, then $\mathcal{O}_\sigma = \mathcal{O}$.

(iii) Every closed set is a G_δ , therefore every F_σ is a $G_{\delta\sigma}$.

(iv) Every open set is an F_σ , therefore every G_δ is an $F_{\sigma\delta}$.

(v) The F_σ 's are the complements of the G_δ 's, and the G_δ 's are the complements of the F_σ 's.

(vi) If E is any Lebesgue-measurable set, then $F \subset E \subset G$ with G a $G_{\delta\sigma}$, F an F_σ and $\lambda(G - F) = 0$.

{Hint: Express E as the union of a sequence (E_n) of measurable sets of finite measure and apply Theorems 2.4.14 and 2.4.18 to each E_n .}

(vii) Iteration of the δ and σ 'operations' can be pursued transfinitely to classify all Borel sets.⁸

5. (H. Steinhaus⁹) If A is a Lebesgue-measurable subset of \mathbb{R} such that $\lambda(A) > 0$, then the difference set $D = \{a - b : a, b \in A\}$ is a neighborhood of 0, that is, $(-\delta, \delta) \subset D$ for some $\delta > 0$.

{Sketch: By 2.4.18, one can suppose that A is compact. For every positive integer n , let

$$U_n = \{x \in \mathbb{R} : |x - a| < 1/n \text{ for some } a \in A\} = \bigcup_{a \in A} \left(a - \frac{1}{n}, a + \frac{1}{n}\right);$$

(U_n) is a decreasing sequence of open sets with $A = \bigcap_{n=1}^{\infty} U_n$, whence $\lambda(A) = \inf \lambda(U_n)$. Choose an index m such that $\lambda(U_m) < \frac{3}{2}\lambda(A)$ and let $\delta = 1/m$. If $|z| < \delta$ then $z \in D$. For, let $B = A + z = \{a + z : a \in A\}$. Note that $B \subset U_m$ (if $x \in B$, say $x = a + z$ with $a \in A$, then $|x - a| = |z| < \delta = 1/m$). Then

$$\begin{aligned} \lambda(U_m - A \cap B) &\leq \lambda(U_m - A) + \lambda(U_m - B) = 2\lambda(U_m) - 2\lambda(A) \\ &< 3\lambda(A) - 2\lambda(A) = \lambda(A) < \frac{2}{3}\lambda(U_m), \end{aligned}$$

therefore $A \cap B \neq \emptyset$ and $a' = a + z$ for suitable $a', a \in A$.}

6. (i) If X is any set, then the power set $\mathcal{P}(X)$ is a commutative ring with unity element X for the operations $A + B = (A - B) \cup (B - A)$ (called the *symmetric difference* of A and B , also denoted $A \Delta B$) and $AB = A \cap B$. Moreover, $AA = A$ for all $A \in \mathcal{P}(X)$ (such rings are called *Boolean*).

(ii) A set $\mathcal{S} \subset \mathcal{P}(X)$ is an algebra of subsets of X if and only if $X \in \mathcal{S}$ and \mathcal{S} is a subring of the ring $\mathcal{P}(X)$ of part (i).

⁸ K. Kuratowski, *Topologie. I* [Monografie Matematyczne, 2nd. edn., Warsaw, 1948], §26, p. 251, II; F. Hausdorff, *Set theory* [3rd. edn., Chelsea, New York, 1957], §18.

⁹ Cf. E. Asplund and L. Bungart, *A first course in integration* [Holt, Rinehart and Winston, New York, 1966], p. 125, Theorem 3.3.9.

7. If \mathcal{E} is a set of subsets of a set X , the σ -algebra $\mathcal{S}(\mathcal{E})$ generated by \mathcal{E} may be obtained by a transfinite procedure.¹⁰ Let Ω be the first uncountable ordinal (1.14.36) and, for each ordinal $\alpha < \Omega$, define \mathcal{E}_α recursively as follows: $\mathcal{E}_0 = \mathcal{E} \cup \{\emptyset, X\}$ and, for $\alpha > 0$,

$$\mathcal{E}_\alpha = \left(\bigcup_{\beta < \alpha} \mathcal{E}_\beta \right)^*,$$

where, for any set $\mathcal{C} \subset \mathcal{P}(X)$, \mathcal{C}^* denotes the set of all countable unions of differences of sets of \mathcal{C} , that is,

$$\mathcal{C}^* = \{ \bigcup_{n=1}^{\infty} (A_n - B_n) : A_n, B_n \in \mathcal{C} \ (n = 1, 2, 3, \dots) \}.$$

(i) $\mathcal{E}_\alpha \subset \mathcal{S}(\mathcal{E})$ for all $\alpha < \Omega$.

(ii) $\alpha < \beta \Rightarrow \mathcal{E}_\alpha \subset \mathcal{E}_\beta$.

(iii) $A, B \in \mathcal{E}_\alpha \Rightarrow A - B \in \mathcal{E}_{\alpha+1}$.

(iv) $A_n \in \mathcal{E}_\alpha \ (n = 1, 2, 3, \dots) \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{E}_{\alpha+1}$.

(v) $\mathcal{S}(\mathcal{E}) = \bigcup_{\alpha < \Omega} \mathcal{E}_\alpha$.

(vi) If $\text{card } \mathcal{E} \leq c$, then $\text{card } \mathcal{S}(\mathcal{E}) \leq c$.

{Hint: (i) Transfinite induction.}

(v) Let \mathcal{S} be the union in question. By (i), $\mathcal{S} \subset \mathcal{S}(\mathcal{E})$. To prove the reverse inclusion, it suffices to show that \mathcal{S} is a σ -algebra. If $\alpha_n < \Omega$ for $n = 1, 2, 3, \dots$, then there exists an $\alpha < \Omega$ such that $\alpha_n < \alpha$ for all n (§1.14, Exercise 12); it follows easily that \mathcal{S} is closed under countable unions.

(vi) Since $\text{card } \{\alpha : \alpha < \Omega\} = \text{card } I(\Omega) = \text{card } \Omega = \aleph_1 \leq c$ (1.14.36), it follows from the formula of (v) that we need only show that $\text{card } \mathcal{E}_\alpha \leq c$ for all $\alpha < \Omega$ (§1.13, Exercise 1). By a transfinite induction, this comes down to showing that $\text{card } \mathcal{C} \leq c \Rightarrow \text{card } \mathcal{C}^* \leq c$, which follows easily from §1.13, Exercise 3.}

8. The Lebesgue measure space $(\mathbb{R}, \mathcal{M}, \lambda)$ of 2.4.13 has the property that if $E \in \mathcal{M}$ and $\lambda(E) = 0$, then every subset of E also belongs to \mathcal{M} and has measure 0 (2.1.7, 2.2.3). A measure space with this property is said to be *complete*.

(i) The measure space $(\mathbb{R}, \mathcal{B}, \lambda|_{\mathcal{B}})$ of 2.4.13 is *not* complete.

{Hint: Exercises 2 and 7.}

(ii) If ρ is an outer measure on a set X , and if \mathcal{S} is the set of all ρ -measurable subsets of X (2.2.7), then the measure space $(X, \mathcal{S}, \rho|_{\mathcal{S}})$ of 2.4.13 is complete.

9. If \mathcal{S} is a σ -algebra of subsets of a set X , a function $\mu : \mathcal{S} \rightarrow [0, +\infty]$ is a measure (in the sense of 2.4.12) if and only if (i) $\mu(\emptyset) = 0$, (ii) $\mu(E \cup F) = \mu(E) + \mu(F)$ when E, F are disjoint sets in \mathcal{S} , and (iii) if $\{F_n\}$ is an

¹⁰ Cf. Hewitt and Stromberg, op. cit., Theorem 10.23; P. R. Halmos, *Measure theory* [Springer-Verlag, New York, 1974], §5, p. 26, Exercise 9.

increasing sequence of sets in \mathcal{S} with union F (that is, $F_n \uparrow F$) then $\mu(F_n) \uparrow \mu(F)$.

{Hint: To see that a measure satisfies (iii), review the proof of 2.2.11. To prove that properties (ii) and (iii) imply countable additivity, let (E_n) be a sequence of pairwise disjoint sets in \mathcal{S} and consider the sequence (F_n) defined by $F_n = \bigcup_{k=1}^n E_k$.}

2.5. A Nonmeasurable Set

Subsets of \mathbb{R} that are not Lebesgue-measurable come into the world reluctantly, requiring the aid of the Axiom of Choice. The reader who doesn't mind an unmotivated proof of the existence of such sets can skip the next paragraphs and go directly to Theorem 2.5.1.

Most constructions of a nonmeasurable set are variations on the following basic idea. One would like to find a set B with $0 < \lambda^*(B) < +\infty$ and a partition $B = \bigcup_{n=1}^{\infty} A_n$ of B into a sequence of pairwise disjoint sets A_n such that $\lambda^*(A_1) = \lambda^*(A_2) = \lambda^*(A_3) = \dots$. It is then easy to see that some A_n must be nonmeasurable; for, if every A_n were measurable then B would be measurable and

$$\lambda(B) = \lambda\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \lambda(A_n) = \sum_{n=1}^{\infty} \lambda(A_1)$$

(a constant term infinite series); since $\lambda(B) > 0$, necessarily $\lambda(A_1) > 0$, whence $\lambda(B) = \sum_{n=1}^{\infty} \lambda(A_1) = +\infty$, a contradiction.

Suppose $0 < \lambda^*(B) < +\infty$ and (A_n) is a sequence of pairwise disjoint nonempty sets such that $B = \bigcup_{n=1}^{\infty} A_n$. A way to assure that the A_n all have the same outer measure is to arrange (if possible) that they are all translates of the same set A , say $A_n = x_n + A$ for all n (cf. 2.1.10). The terms of the sequence (x_n) will then be distinct (because the A_n are pairwise disjoint) and, writing $D = \{x_n : n \in \mathbb{P}\}$, we have $B = \bigcup_{n=1}^{\infty} (x_n + A) = D + A$. Indeed, every $b \in B$ has a *unique* representation $b = x + a$ with $x \in D$ and $a \in A$. {If $b = x_n + a = x_m + a'$ with $a, a' \in A$, then $b \in x_n + A = A_n$ and $b \in x_m + A = A_m$, $b \in A_n \cap A_m$, whence $n = m$, $x_n = x_m$ and $a = a'$.} Since $\lambda^*(B) > 0$, B is uncountable (2.1.3, (iii)); but D is countable, so A must be uncountable. The 'direct sum' $B = D + A = A + D$ effects a partition of B into countably many uncountable sets,

$$B = \bigcup_{x \in D} (x + A),$$

as well as a partition of B into uncountably many countable sets,

$$B = \bigcup_{a \in A} (a + D).$$

Translates $a + \mathbb{Q}$ remind us of cosets, and \mathbb{Q} is a countable subgroup of the additive group \mathbb{R} .

We are thus led to consider the quotient (additive) group \mathbb{R}/\mathbb{Q} and the partition of \mathbb{R} into cosets $x + \mathbb{Q}$. Since \mathbb{Q} is countable and \mathbb{R} is not (1.10.10, 1.10.11), there are uncountably many cosets. From each coset u choose one representative a_u and let $A = \{a_u : u \in \mathbb{R}/\mathbb{Q}\}$ be the resulting set of representatives. {For example, let $\gamma : \mathcal{P}(\mathbb{R}) - \{\emptyset\} \rightarrow \mathbb{R}$ be a choice function for \mathbb{R} (1.4.6) and let $a_u = \gamma(u)$ for all $u \in \mathbb{R}/\mathbb{Q}$.} Thus, A is a subset of \mathbb{R} such that, for every coset $x + \mathbb{Q}$, $A \cap (x + \mathbb{Q})$ is a singleton. Then

$$(*) \quad \mathbb{R} = A + \mathbb{Q}.$$

For, every real number b belongs to some coset u (namely, $u = b + \mathbb{Q}$), whence $b - a_u \in \mathbb{Q}$, say $b - a_u = r \in \mathbb{Q}$; then $b = a_u + r \in A + \mathbb{Q}$. Moreover, representations in $(*)$ are unique: if $a + r = a' + r'$ with $a, a' \in A$ and $r, r' \in \mathbb{Q}$, then $a - a' = r' - r \in \mathbb{Q}$, so a and a' belong to the same coset, whence $a = a'$ and $r = r'$. Let $\mathbb{Q} = \{r_n : n \in \mathbb{P}\}$ be a faithful indexing of \mathbb{Q} by \mathbb{P} (1.10.10). We have

$$\mathbb{R} = \mathbb{Q} + A = \bigcup_{n \in \mathbb{P}} (r_n + A);$$

writing $A_n = r_n + A$, we have a denumerable partition $\mathbb{R} = \bigcup_{n=1}^{\infty} A_n$ with $\lambda^*(A_n)$ constant. The only catch is that $\lambda(\mathbb{R}) = +\infty$, so \mathbb{R} cannot play the role of B in the earlier discussion; nevertheless, it can be shown that A is nonmeasurable (Exercise 4), but it is simpler to play a 'bounded variation' on the preceding construction:

2.5.1. Theorem (G. Vitali¹). *There exists a subset of \mathbb{R} that is not Lebesgue-measurable.*

*Proof.*² For $x, y \in (0, 1)$, write $x \sim y$ if $x - y$ is rational; this defines an equivalence relation in $(0, 1)$. Note that

$$\{x - y : x, y \in (0, 1)\} = (-1, 1),$$

thus $x \sim y \Leftrightarrow x - y \in \mathbb{Q} \cap (-1, 1)$. Let $n \mapsto r_n$ be a faithful indexing of the denumerable set $D = \mathbb{Q} \cap (-1, 1)$.

By the Axiom of Choice, there exists a subset A of $(0, 1)$ such that A intersects each equivalence class in a singleton; let

$$B = D + A = \{r_n + a : n \in \mathbb{P}, a \in A\}.$$

Writing $A_n = r_n + A$, we have a denumerable partition $B = \bigcup_{n=1}^{\infty} A_n$ with $\lambda^*(A_n) = \lambda^*(A)$ for all n (2.1.10).

¹ Giuseppe Vitali (1875-1932)

² Patterned after the exposition of H. Kestelman, *Modern theories of integration* [Oxford Univ. Press, 1937; 2nd. revised edn., Dover, New York, 1960], pp. 90-91.

Note that $(0, 1) \subset B$; for, if $x \in (0, 1)$ and a is the element of A that belongs to the equivalence class of x , then $x \sim a$, thus $x - a \in D$ and $x \in D + A = B$. On the other hand, since $A \subset (0, 1)$ and $D \subset (-1, 1)$, we have $B = A + D \subset (-1, 2)$. Thus

$$(*) \quad (0, 1) \subset B \subset (-1, 2),$$

whence $1 \leq \lambda^*(B) \leq 3$.

Finally, we assert that A is not Lebesgue-measurable. If, on the contrary, A were measurable, then the sets $A_n = r_n + A$ and $B = \bigcup_{n=1}^{\infty} A_n$ would be measurable, and, by the countable additivity of λ ,

$$\lambda(B) = \sum_{n=1}^{\infty} \lambda(A_n) = \sum_{n=1}^{\infty} \lambda(A).$$

Since $\lambda(B) > 0$, necessarily $\lambda(A) > 0$; but then $\lambda(B) = \sum_{n=1}^{\infty} \lambda(A) = +\infty$, a contradiction. \diamond

Exercises

1. (i) If S is a subset of \mathbb{R} such that, for every $A \subset S$,

$$(*) \quad \lambda^*(A) = \sup\{\lambda(K) : K \subset A, K \text{ compact}\},$$

then S is Lebesgue-measurable.

(ii) If S is a nonmeasurable set (2.5.1), then $(*)$ must fail for some subset A of S .

{Hint: (i) We can suppose without loss of generality that $\lambda^*(S) < +\infty$. Show that there exists a Borel subset E of S (in fact, the union of a sequence of compact subsets) having the same outer measure as S , then repeat the argument to obtain a Borel subset F of $S - E$ with the same outer measure as $S - E$. Apply λ^* to the inclusion $S \supset E \cup F$ to conclude that $S - E$ has outer measure 0.}

2. Let $B = \bigcup_{n=1}^{\infty} A_n$ as in the proof of Theorem 2.5.1.

(i) There exists a positive integer n such that $\lambda^*(\bigcup_{i=1}^n A_i) < \sum_{i=1}^n \lambda^*(A_i)$.

(ii) If m is the smallest positive integer such that $\lambda^*(\bigcup_{i=1}^m A_i) < \sum_{i=1}^m \lambda^*(A_i)$ and if $S = \bigcup_{i=1}^{m-1} A_i$, $T = A_m$, then $S \cap T = \emptyset$ and $\lambda^*(S \cup T) < \lambda^*(S) + \lambda^*(T)$.

(iii) The set S of part (ii) is nonmeasurable.

{Hints: (i) Consider the alternative.

(ii) Argue that $\lambda^*(\bigcup_{i=1}^{m-1} A_i) = \sum_{i=1}^{m-1} \lambda^*(A_i)$.

(iii) Use (ii) to show that S does not split $S \cup T$ additively.}

3. There are 2^c nonmeasurable subsets of \mathbb{R} .

4. The set A with $\mathbb{R} = A + \mathbb{Q}$ described in the remarks preceding 2.5.1 is nonmeasurable.

{Hint: Since $\mathbb{R} = \bigcup_{r \in \mathbb{Q}} (r + A)$ and $\lambda^*(r + A) = \lambda^*(A)$ for all $r \in \mathbb{Q}$, necessarily $\lambda^*(A) > 0$. Let $D = \{a - b : a, b \in A\}$; since $D \cap \mathbb{Q} = \{0\}$ by the choice of A , D can have no interior points, therefore A is nonmeasurable (§2.4, Exercise 5).}

2.6. Abstract Measure Spaces

We note in this section a number of useful properties of Lebesgue measure that carry over to the general measure spaces defined in 2.4.12. Throughout the section, (X, \mathcal{S}, μ) is a fixed measure space.

2.6.1. Theorem. (1) μ is **finitely additive** on \mathcal{S} , that is, if E_1, \dots, E_r are pairwise disjoint sets in \mathcal{S} , then $\mu(\bigcup_{k=1}^r E_k) = \sum_{k=1}^r \mu(E_k)$.

(2) μ is **monotone**, that is, if E, F are sets in \mathcal{S} with $E \subset F$, then $\mu(E) \leq \mu(F)$. If, moreover, $\mu(F) < +\infty$, then $\mu(F - E) = \mu(F) - \mu(E)$.

Proof. (1) Write $E_1 \cup \dots \cup E_r = E_1 \cup \dots \cup E_r \cup \emptyset \cup \emptyset \cup \dots$ and cite the countable additivity of μ and the property $\mu(\emptyset) = 0$.

(2) If $E \subset F$ then $F = E \cup (F - E)$ is a disjoint union; by finite additivity, $\mu(F) = \mu(E) + \mu(F - E) \geq \mu(E)$; when $\mu(F)$ is finite, the term $\mu(E)$ can be transposed in the equality. \diamond

2.6.2. Corollary. If F and F_n ($n = 1, 2, 3, \dots$) are sets in \mathcal{S} such that $F_n \uparrow F$, then $\mu(F_n) \uparrow \mu(F)$.

Proof. A simplification of the proof of 2.2.11, with A replaced by X and λ^* by μ . \diamond

2.6.3. Corollary. If G and G_n ($n = 1, 2, 3, \dots$) are sets in \mathcal{S} such that $G_n \downarrow G$ and if $\mu(G_1) < +\infty$, then $\mu(G_n) \downarrow \mu(G)$.

Proof. A simplification of the proof of 2.2.12. \diamond

2.6.4. Corollary. For every sequence (E_n) in \mathcal{S} , $\mu(\bigcup_{k=1}^{\infty} E_k) \leq \sum_{k=1}^{\infty} \mu(E_k)$.

Proof. Writing $F_n = \bigcup_{k=1}^n E_k$ and $F = \bigcup_{k=1}^{\infty} E_k$, we have $\mu(F_n) \uparrow \mu(F)$ by 2.6.2. Since

$$F_2 = E_1 \cup E_2 = (E_1 - E_2) \cup (E_1 \cap E_2)$$

and the terms of the union on the right are disjoint, we have

$$\mu(F_2) = \mu(E_1 - E_2) + \mu(E_1 \cap E_2) \leq \mu(E_1) + \mu(E_2);$$

by an obvious induction,

$$\mu(F_n) \leq \sum_{k=1}^n \mu(E_k) \leq \sum_{k=1}^{\infty} \mu(E_k)$$

for all n , whence $\mu(F) = \sup \mu(F_n) \leq \sum_{k=1}^{\infty} \mu(E_k)$. \diamond

Exercises

1. For a sequence (E_n) of subsets of X , one defines

$$\begin{aligned} \limsup E_n &= \{x \in X : x \in E_n \text{ frequently}\}, \\ \liminf E_n &= \{x \in X : x \in E_n \text{ ultimately}\}. \end{aligned}$$

(i) Show that

$$\begin{aligned} \limsup E_n &= \bigcap_{n=1}^{\infty} \left(\bigcup_{k=n}^{\infty} E_k \right), \\ \liminf E_n &= \bigcup_{n=1}^{\infty} \left(\bigcap_{k=n}^{\infty} E_k \right). \end{aligned}$$

In particular, if the E_n are in \mathcal{S} then so are $\limsup E_n$ and $\liminf E_n$.

(ii) If the E_n belong to \mathcal{S} then

$$\mu(\liminf E_n) \leq \liminf \mu(E_n)$$

(for the meaning of the right side, see §1.16).

(iii) Suppose the E_n belong to \mathcal{S} and $\mu(\bigcup_{k=1}^{\infty} E_k) < +\infty$. Then

$$\mu(\limsup E_n) \geq \limsup \mu(E_n);$$

in particular, if $\mu(E_n) \geq r > 0$ for all n then $\mu(\limsup E_n) \geq r$ (*Arzela-Young theorem*).

2. If $E \subset F$ and $\mu(E)$ is finite, then $\mu(F - E) = \mu(F) - \mu(E)$.

3. In the definition of a measure (2.4.12), the condition (2) that μ is countably additive may be replaced by the following pair of conditions:

(2a) μ is finitely additive, and

(2b) if (F_n) is an increasing sequence of sets in \mathcal{S} with union F , then $\mu(F_n) \uparrow \mu(F)$ in $\overline{\mathbb{R}}$, that is, $(\mu(F_n))$ is an increasing sequence with supremum $\mu(F)$.

{Hint: Assuming (2a) and (2b), and given a sequence of pairwise disjoint sets E_n in \mathcal{S} , contemplate the sequence $F_n = \bigcup_{k=1}^n E_k$ with union $F = \bigcup_{n=1}^{\infty} E_n$.}

4. If \mathcal{S} is a σ -algebra of subsets of a set X , a set function $\mu : \mathcal{S} \rightarrow [0, +\infty]$ is a measure if and only if (i) $\mu(\emptyset) = 0$, (ii) $\mu(E \cup F) = \mu(E) + \mu(F)$ when E, F are disjoint sets in \mathcal{S} , and (iii) $\mu(\bigcup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} \mu(E_n)$ for every sequence (E_n) of pairwise disjoint sets in \mathcal{S} .

{Hint: "If": Assuming that μ satisfies (i)–(iii), argue that μ is monotone and finitely additive, then infer the reverse of the inequality in (iii) for a sequence of pairwise disjoint sets in \mathcal{S} .}

CHAPTER 3

Topology

§3.1. Metric spaces: examples

§3.2. Convergence, closed sets and open sets in metric spaces

§3.3. Topological spaces

§3.4. Continuity

§3.5. Limit of a function

The following informal remarks are intended as motivation for the subject of this chapter; the reader who already knows what a topological space is may prefer to plunge right into §3.1.

The usual definition of sequential convergence $a_n \rightarrow a$ in \mathbb{R} is as follows: for every $\delta > 0$ there is an index N such that $n \geq N \Rightarrow |a_n - a| < \delta$. The parameters of this definition contain not only qualitative information (that ‘ a_n approaches a as n tends to infinity’) but also quantitative information: if, for each $\delta > 0$, N_δ is the smallest positive integer such that $n \geq N_\delta \Rightarrow |a_n - a| < \delta$, then the function $\delta \mapsto N_\delta$ contains information about the *rate* at which a_n converges to a .

The *continuity* of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ at a point $a \in \mathbb{R}$ can be expressed by saying that $a_n \rightarrow a \Rightarrow f(a_n) \rightarrow f(a)$. Already, the quantitative information just mentioned has lost some of its impact, since the rate of convergence of $f(a_n)$ to $f(a)$ (which depends also on the function f) has also entered the picture. A quantitative handle on the continuity of f at a is suggested by expressing continuity in terms of ϵ ’s and δ ’s: for each $\epsilon > 0$ there exists a $\delta > 0$ such that

$$(*) \quad |x - a| < \delta \Rightarrow |f(x) - f(a)| < \epsilon.$$

If, for each $\epsilon > 0$, we define δ_ϵ to be the supremum of all $\delta > 0$ for which the implication $(*)$ is valid (conceivably $\delta_\epsilon = +\infty$ —consider, for example, a constant function), then δ_ϵ is the largest $\delta \in (0, +\infty]$ satisfying $(*)$. The function $\epsilon \mapsto \delta_\epsilon$ conveys numerical information that is more pertinent to the *rate of change* of f than to its continuity. Finally, when we contemplate functions f that are continuous at *every* point a of \mathbb{R} , we are dealing with a function $(a, \epsilon) \mapsto \delta_{a, \epsilon}$ ($a \in \mathbb{R}$, $\epsilon > 0$) of two variables; clearly there is a lot of numerical baggage here that goes beyond the intuitive concept of continuity.

Here is a way of filtering out some of this numerical 'static'. To say that $a_n \rightarrow a$ means that for every open interval I containing a , $a_n \in I$ ultimately. To say that f is continuous at a means that if J is any open interval containing $f(a)$, then there exists an open interval I containing a such that $f(I) \subset J$; or, relaxing the notations a bit, if W is any set that contains an open interval containing $f(a)$, then $f^{-1}(W)$ contains an open interval containing a . In discussing functions f that are continuous at every point of \mathbb{R} , one is naturally led to consider subsets U of \mathbb{R} such that, for every $x \in U$, U contains an open interval containing x ; such subsets of \mathbb{R} are called *open sets* and open sets are the basis of the concept of topological space.

General topological spaces are taken up in §3.3; in §§3.1, 3.2, we prepare the ground for the concept by looking at the precursors of the topological ideas in the context of metric spaces (a special kind of topological space in which there is assigned a numerical measure of 'distance' between any two points of the space).

3.1. Metric Spaces: Examples

Metric spaces provide a unified setting for discussing a multitude of sequential convergence matters. Our discussion of examples of metric spaces rests on several inequalities, the first mainly of technical interest, the rest eminently practical.

3.1.1. Proposition. If a, b, c are real numbers ≥ 0 such that $a \leq b + c$, then

$$\frac{a}{1+a} \leq \frac{b}{1+b} + \frac{c}{1+c}.$$

Proof. Elementary. {Incidentally, the converse is false: consider $b = c = 1$ and $a = 3$.} \diamond

3.1.2. Proposition. For any two ordered pairs $(a, b), (c, d)$ of real numbers,

- (1) $ac + bd \leq (a^2 + b^2)^{1/2}(c^2 + d^2)^{1/2}$,
- (2) $[(a + c)^2 + (b + d)^2]^{1/2} \leq (a^2 + b^2)^{1/2} + (c^2 + d^2)^{1/2}$.

Proof. Write $R = (a^2 + b^2)^{1/2}$, $S = (c^2 + d^2)^{1/2}$. Squaring both sides of (1) and subtracting,

$$R^2 S^2 - (ac + bd)^2 = (ad - bc)^2 \geq 0,$$

whence the inequality (1); the identity

$$(R + S)^2 - [(a + c)^2 + (b + d)^2] = 2[RS - (ac + bd)]$$

shows that the inequalities of (1) and (2) are equivalent. \diamond

3.1.3. Proposition. If p and q are positive real numbers such that

$$\frac{1}{p} + \frac{1}{q} = 1$$

(in particular, $p > 1$ and $q > 1$), then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

for all nonnegative real numbers a and b .

Proof. We can suppose that $a > 0$ and $b > 0$. Consider the function $\varphi : (0, +\infty) \rightarrow (0, +\infty)$ defined by

$$\varphi(t) = \frac{t^p}{p} + \frac{t^{-q}}{q} \quad (t > 0);$$

φ is differentiable (therefore continuous), with $\varphi'(t) = t^{p-1} - t^{-q-1}$, whence $\varphi'(1) = 0$ and φ' has no other zeros. The function φ has a minimum at $t = 1$; this follows from the fact that $\varphi(t) \rightarrow +\infty$ as $t \downarrow 0$ or $t \uparrow +\infty$, while $t = 1$ is the only possible local extremum (max or min) of φ . (A more explicit argument is sketched in Exercise 1.)

Summarizing, we have shown that

$$(*) \quad 1 \leq \frac{t^p}{p} + \frac{t^{-q}}{q} \quad \text{for all } t > 0.$$

Substituting $t = a^{1/q}b^{-1/p}$ in $(*)$, we have

$$1 \leq \frac{a^{p/q}b^{-1}}{p} + \frac{a^{-1}b^{q/p}}{q},$$

and multiplication by ab yields

$$(**) \quad ab \leq \frac{a^{1+p/q}}{p} + \frac{b^{1+q/p}}{q};$$

but

$$1 + \frac{p}{q} = p \left(\frac{1}{p} + \frac{1}{q} \right) = p$$

and similarly $1 + q/p = q$, thus $(**)$ simplifies to the desired inequality. \diamond

The importance of inequalities for metric spaces is built right into the definition:

3.1.4. Definition. Let X be a nonempty set. A **metric** on X is a function

$$d : X \times X \rightarrow [0, +\infty)$$

assigning, to each ordered pair (x, y) of points of X , a nonnegative real number $d(x, y)$, such that

- (i) $d(x, y) \leq d(x, z) + d(z, y)$ for all x, y, z in X (*triangle inequality*),
- (ii) $d(x, y) = d(y, x)$ for all x, y in X (*symmetry*),
- (iii) $d(x, x) = 0$ for all $x \in X$,
- (iv) if $x \neq y$ then $d(x, y) > 0$ (*strict positivity*).

If condition (iv) is omitted, d is called a *pseudometric* on X .

A **metric space** (or pseudometric space) is a pair (X, d) , where d is a metric (or pseudometric) on X . One calls d the *distance function* of the space, and $d(x, y)$ the *distance from x to y* .

3.1.5. *Remark.* If d is a pseudometric on X then the relation $x \sim y$ in X defined by $d(x, y) = 0$ is an equivalence relation; writing $\tilde{x} = \{t : d(t, x) = 0\}$ for the equivalence class of x , the formula $\tilde{d}(\tilde{x}, \tilde{y}) = d(x, y)$ defines a metric \tilde{d} on the quotient set. {To see that $d(x, y)$ depends only on \tilde{x} and \tilde{y} , note that if $\tilde{x} = \tilde{s}$ and $\tilde{y} = \tilde{t}$, then $d(x, y) \leq d(x, s) + d(s, t) + d(t, y) = d(s, t)$.}

A definition is worth its weight in examples:

3.1.6. *Example.* $X = \mathbb{R}$ and $d(x, y) = |x - y|$ (the 'usual metric' on \mathbb{R}). The triangle inequality reduces to the inequality $|a + b| \leq |a| + |b|$ via the identity $x - y = (x - z) + (z - y)$.

3.1.7. *Example.* X any nonempty set and

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y. \end{cases}$$

One calls d the *discrete metric* on X , and (X, d) a *discrete metric space*.

3.1.8. *Example.* If d is a metric (or a pseudometric) on a set X , then so is the function D defined by

$$D(x, y) = \frac{d(x, y)}{1 + d(x, y)}$$

(the triangle inequality for D follows from 3.1.1). Moreover, $0 \leq D(x, y) < 1$ and

$$d(x, y) = \frac{D(x, y)}{1 - D(x, y)}.$$

3.1.9. *Example.* Let $X = \mathbb{C}$ be the set of all complex numbers $z = x + iy$ ($x, y \in \mathbb{R}$). If $\bar{z} = x - iy$ is the conjugate of $z = x + iy$, then the absolute value of z is defined by $|z| = (z\bar{z})^{1/2} = (x^2 + y^2)^{1/2}$. The formula $d(z, z') = |z - z'|$ defines a metric on \mathbb{C} , the triangle inequality being immediate from (2) of 3.1.2.

3.1.10. *Example.* Let T be a nonempty set. A complex-valued function $x : T \rightarrow \mathbb{C}$ is said to be *bounded* if

$$\sup_{t \in T} |x(t)| < +\infty,$$

that is, the set $\{|x(t)| : t \in T\}$ is a bounded subset of \mathbb{R} . We write $\mathcal{B} = \mathcal{B}(T, \mathbb{C}) = \mathcal{B}_{\mathbb{C}}(T)$ for the set of all such x ; \mathcal{B} is a complex vector space for the pointwise linear operations, defined by the formulas

$$(x + y)(t) = x(t) + y(t), \quad (cx)(t) = cx(t).$$

If $x \in \mathcal{B}$ the nonnegative real number $\|x\|_{\infty}$ defined by the formula

$$\|x\|_{\infty} = \sup_{t \in T} |x(t)|$$

is called the *sup-norm* of x . If $x, y \in \mathcal{B}$ then, for all $t \in T$,

$$|(x + y)(t)| = |x(t) + y(t)| \leq |x(t)| + |y(t)| \leq \|x\|_{\infty} + \|y\|_{\infty}$$

by the triangle inequality for \mathbb{C} (3.1.9), therefore $\|x + y\|_{\infty} \leq \|x\|_{\infty} + \|y\|_{\infty}$. It follows that the formula

$$d(x, y) = \|x - y\|_{\infty} = \sup_{t \in T} |x(t) - y(t)|$$

defines a metric d on \mathcal{B} , called the *sup-metric*, also denoted d_{∞} . Several other important examples can be obtained by specialization:

(i) The set $\mathcal{B}(T, \mathbb{R}) = \mathcal{B}_{\mathbb{R}}(T)$ of all bounded functions $x : T \rightarrow \mathbb{R}$ is a real-linear subspace of $\mathcal{B}(T, \mathbb{C})$; it, too, is a metric space for the sup-metric.

(ii) If $T = \{1, 2, \dots, n\}$, then $\mathcal{B}(T, \mathbb{C})$ may be identified with the complex vector space \mathbb{C}^n of all n -ples of complex numbers, the function x corresponding to an n -ple (x_1, x_2, \dots, x_n) being defined by $x(k) = x_k$ for $k = 1, \dots, n$. Thus \mathbb{C}^n is a metric space for the sup-metric

$$d_{\infty}(x, y) = \max\{|x_k - y_k| : k = 1, \dots, n\},$$

as is the set \mathbb{R}^n of n -ples of real numbers. (The reason for the subscript ∞ is explained in Exercise 3.)

The next series of examples depends on two inequalities derived from 3.1.3. First, a definition:

3.1.11. *Definition.* If $x = (x_1, \dots, x_n) \in \mathbb{C}^n$ and p is a real number ≥ 1 , the nonnegative real number

$$\|x\|_p = \left(\sum_{k=1}^n |x_k|^p \right)^{1/p}$$

is called the (Minkowski) **p -norm** of x .

Of course if p happens to be a positive integer ($p = 1$ and $p = 2$ being the most important examples), then $1/p$ indicates p 'th root.¹ When $p = 2$ and the coordinates of x are real, $\|x\|_2$ is called the *Euclidean norm* of the vector x (with the usual interpretation as 'length' when $n = 2$ or $n = 3$). In general, it is clear that $\|cx\|_p = |c| \|x\|_p$ for all complex numbers c , and $\|x\|_p = 0$ if and only if x is the zero vector.

3.1.12. Lemma. (*Hölder's inequality*) If $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ are n -ples of complex numbers and if $\frac{1}{p} + \frac{1}{q} = 1$ as in 3.1.3, then

$$\left| \sum_{k=1}^n x_k y_k \right| \leq \|x\|_p \|y\|_q.$$

Proof. By the triangle inequality in \mathbb{C} , we have

$$\left| \sum_{k=1}^n x_k y_k \right| \leq \sum_{k=1}^n |x_k y_k| = \sum_{k=1}^n |x_k| |y_k|,$$

so we can suppose that the x_k, y_k are nonnegative real numbers. Let $a = \|x\|_p$, $b = \|y\|_q$; we are to show that $|\sum_{k=1}^n x_k y_k| \leq ab$. If $a = 0$ then x is the zero vector and the inequality is trivial. Similarly if $b = 0$. Suppose that $a > 0$ and $b > 0$. Replacing x and y by $a^{-1}x$ and $b^{-1}y$, we can suppose that $\|x\|_p = \|y\|_q = 1$, in other words,

$$\sum_{k=1}^n (x_k)^p = \sum_{k=1}^n (y_k)^q = 1;$$

the problem is to show that $\sum_{k=1}^n x_k y_k \leq 1$. For $k = 1, \dots, n$ we have

$$x_k y_k \leq \frac{(x_k)^p}{p} + \frac{(y_k)^q}{q}$$

by 3.1.3, therefore

$$\sum_{k=1}^n x_k y_k \leq \frac{\sum_{k=1}^n (x_k)^p}{p} + \frac{\sum_{k=1}^n (y_k)^q}{q} = \frac{1}{p} + \frac{1}{q} = 1. \quad \diamond$$

3.1.13. Proposition. (*Minkowski's inequality*) With notations as in the lemma, $\|x + y\|_p \leq \|x\|_p + \|y\|_p$.

Proof. We can suppose that $x + y$ is not the zero vector; writing $w_k = |x_k + y_k|$ for $k = 1, \dots, n$, we have

$$\sum_{k=1}^n (w_k)^p = (\|x + y\|_p)^p > 0.$$

¹ First course, p. 159, 9.5.15.

For each k ,

$$(w_k)^p = |x_k + y_k|(w_k)^{p-1} \leq |x_k|(w_k)^{p-1} + |y_k|(w_k)^{p-1};$$

summing on k and citing the lemma,

$$\begin{aligned} \sum_{k=1}^n (w_k)^p &\leq \sum_{k=1}^n |x_k|(w_k)^{p-1} + \sum_{k=1}^n |y_k|(w_k)^{p-1} \\ &\leq \left(\sum_{k=1}^n |x_k|^p \right)^{\frac{1}{p}} \left(\sum_{k=1}^n (w_k)^{(p-1)q} \right)^{\frac{1}{q}} \\ &\quad + \left(\sum_{k=1}^n |y_k|^p \right)^{\frac{1}{p}} \left(\sum_{k=1}^n (w_k)^{(p-1)q} \right)^{\frac{1}{q}} \\ &= (\|x\|_p + \|y\|_p) \left(\sum_{k=1}^n (w_k)^{(p-1)q} \right)^{\frac{1}{q}}. \end{aligned}$$

Since $(p-1)q = p$, the preceding inequality may be written

$$(\|x + y\|_p)^p \leq (\|x\|_p + \|y\|_p)(\|x + y\|_p)^{p/q},$$

therefore

$$(\|x + y\|_p)^{p-p/q} \leq \|x\|_p + \|y\|_p,$$

and the observation $p - p/q = p(1 - 1/q) = p(1/p) = 1$ completes the proof. \diamond

Minkowski's inequality underlies a host of useful metric spaces, especially in integration theory (the L^p -spaces discussed in Chapter 6); for the present, we limit ourselves to a class of examples accessible by elementary means:

3.1.14. *Example.* Let T be a nonempty set, p a real number ≥ 1 , and let X be the set of all functions $x : T \rightarrow \mathbb{C}$ such that

$$\sum_{t \in T} |x(t)|^p < +\infty$$

(cf. 1.15.5). For such a function x , define

$$\|x\|_p = \left(\sum_{t \in T} |x(t)|^p \right)^{1/p};$$

note that $cx \in X$ for all complex numbers c , with $\|cx\|_p = |c| \|x\|_p$, and that $\|x\|_p > 0$ unless x is the function identically zero.

If $x, y \in X$ then $x + y \in X$ and $\|x + y\|_p \leq \|x\|_p + \|y\|_p$. Indeed, for every finite subset F of T ,

$$\left(\sum_{t \in F} |x(t) + y(t)|^p \right)^{1/p} \leq \left(\sum_{t \in F} |x(t)|^p \right)^{1/p} + \left(\sum_{t \in F} |y(t)|^p \right)^{1/p} \leq \|x\|_p + \|y\|_p$$

(by Minkowski's inequality when $p > 1$ and by the triangle inequality in \mathbb{C} when $p = 1$); this shows that $x + y \in X$ and the desired inequality follows on taking supremum over all F . The formula

$$d_p(x, y) = \|x - y\|_p$$

thus defines a metric d_p on X . The metric space (X, d_p) is denoted $l_{\mathbb{C}}^p(T)$, and d_p is called the l^p -metric. Since $x \in l_{\mathbb{C}}^p(T) \Leftrightarrow |x|^p \in l_{\mathbb{C}}^1(T)$, where $|x|^p(t) = |x(t)|^p$ for all $t \in T$, such a function x is said to be p -th power summable; it is zero for all but countably many points of T (1.15.9).

The set of all $x \in l_{\mathbb{C}}^p(T)$ that are real-valued is denoted $l_{\mathbb{R}}^p(T)$. When $T = \{1, 2, \dots, n\}$, the metric

$$d_p(x, y) = \left(\sum_{k=1}^n |x_k - y_k|^p \right)^{1/p}$$

on \mathbb{C}^n is called the *Minkowski p -metric*, and the metric space (\mathbb{C}^n, d_p) is called an n -dimensional complex *Minkowski space*. Similarly, (\mathbb{R}^n, d_p) is called an n -dimension real Minkowski space.

The symbol (l^p) is traditionally reserved for $l_{\mathbb{R}}^p(\mathbb{P})$ (or $l_{\mathbb{C}}^p(\mathbb{P})$), the space of p -th power summable sequences of real (or complex) numbers.

3.1.15. Example. When $p = 2$ in the preceding example, (\mathbb{R}^n, d_2) is called the n -dimensional *Euclidean space* and d_2 the *Euclidean metric* on \mathbb{R}^n , whereas (\mathbb{C}^n, d_2) is called the n -dimensional *unitary space*. For $p = 2$ there are proofs of the triangle inequality simpler than the proof via Hölder's inequality (consult any book on linear algebra).

Exercises

1. (i) The equation $2^x = x$ has no solution, therefore $2^x > x$ for all $x \in \mathbb{R}$.

(ii) If φ is the function of 3.1.3, then $\varphi(t) > 1$ for all $t > 0$ outside the closed interval $[1/2, 2]$. {Hint: If $t > 2$ then $t^p > 2^p > p$.}

2. If p and q are integers such that $\frac{1}{p} + \frac{1}{q} = 1$, then $p = q = 2$.

3. With $\|x\|_\infty$ and $\|x\|_p$ ($p \geq 1$) defined for $x \in \mathbb{C}^n$ as in 3.1.10 and 3.1.11,

$$\|x\|_\infty = \lim_{p \rightarrow \infty} \|x\|_p.$$

{Hint: One can suppose $\|x\|_\infty = 1$. Let m be the number of coordinates x_k of x such that $|x_k| = 1$. Then $(\|x\|_p)^p = m + t_p$, where $t_p \rightarrow 0$ as $p \rightarrow \infty$.}

4. With notations as in 3.1.14, if $1 \leq p < r$ then

$$l_{\mathbb{C}}^p(T) \subset l_{\mathbb{C}}^r(T).$$

{Hint: If $x \in l_{\mathbb{C}}^p(T)$ then $|x(t)| < 1$ for all but finitely many t .}

5. If $a = (a_1, \dots, a_n)$, $b = (b_1, \dots, b_n)$, $c = (c_1, \dots, c_n)$ are elements of \mathbb{C}^n such that $|a_k| \leq |b_k| + |c_k|$ for all k , then $\|a\|_p \leq \|b\|_p + \|c\|_p$ for all $p \geq 1$.

6. Let (X_k, d_k) ($k = 1, 2, \dots, n$) be pseudometric spaces and let $X = X_1 \times \dots \times X_n$ be the product set. For every pair of points $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$ of X , and for every real number $p \geq 1$, each of the formulas

$$d(x, y) = \left(\sum_{k=1}^n [d_k(x_k, y_k)]^p \right)^{1/p}$$

$$d(x, y) = \max_{1 \leq k \leq n} d_k(x_k, y_k)$$

defines a pseudometric on X . In order that these be metrics on X , it is necessary and sufficient that the pseudometrics d_1, \dots, d_n be *separating* in the following sense: if $x, y \in X$ and $x \neq y$, then $d_k(x_k, y_k) > 0$ for some k .

7. If (X_k, d_k) ($k = 1, 2, 3, \dots$) is a sequence of pseudometric spaces and $X = \prod_{k=1}^{\infty} X_k$ is the product set, then the formula

$$d(x, y) = \sum_{k=1}^{\infty} \frac{1}{2^k} \cdot \frac{d_k(x_k, y_k)}{1 + d_k(x_k, y_k)},$$

where $x = (x_1, x_2, x_3, \dots)$, $y = (y_1, y_2, y_3, \dots)$ are points of X , defines a pseudometric on X . For d to be a metric, it is necessary and sufficient that the sequence d_1, d_2, d_3, \dots be 'separating' (in the sense suggested by Exercise 6).

3.2. Convergence, Closed Sets and Open Sets in Metric Spaces

The actual values of a metric can be important in geometrical situations, but what often counts most in analysis is the notion of limit derived

from the metric. In a metric space, the core concept of limit is sequential convergence:

3.2.1. *Definition.* Let (X, d) be a metric space (3.1.4). A sequence (x_n) in X is said to be **convergent** (for d) if there exists a point $x \in X$ such that $d(x_n, x) \rightarrow 0$ as $n \rightarrow \infty$ (that is, for every $\epsilon > 0$, $d(x_n, x) < \epsilon$ ultimately). Such a point x is then unique; for, if also $d(x_n, y) \rightarrow 0$ then

$$0 \leq d(x, y) \leq d(x, x_n) + d(x_n, y) \rightarrow 0$$

shows that $x = y$. One says that x_n **converges** to the **limit** x (in X) and one writes

$$x = \lim_{n \rightarrow \infty} x_n,$$

or $x_n \rightarrow x$ as $n \rightarrow \infty$, or simply $x_n \rightarrow x$.

Note that if (x_n) converges to x , then every subsequence (x_{n_k}) also converges to x .

3.2.2. *Examples.* (i) If $X = \mathbb{R}$ and d is the usual metric on \mathbb{R} (3.1.6), then the concept defined in 3.2.1 is the usual concept of convergence in \mathbb{R} .

(ii) In a discrete metric space (3.1.7),

$$x_n \rightarrow x \Leftrightarrow x_n = x \text{ ultimately};$$

for, if $0 < \epsilon \leq 1$ then 0 is the only available distance $< \epsilon$. Thus the only convergent sequences are the sequences that are ultimately constant.

(iii) If d is a pseudometric on X and if $D = d/(1 + d)$ (cf. 3.1.8), then

$$d(x_n, x) \rightarrow 0 \Leftrightarrow D(x_n, x) \rightarrow 0.$$

In particular, when d is a metric,

$$x_n \rightarrow x \text{ for } d \Leftrightarrow x_n \rightarrow x \text{ for } D.$$

(iv) In \mathbb{C}^r with the sup-metric d_∞ (3.1.10), a sequence is convergent if and only if, for each $k = 1, \dots, r$, the sequence of k 'th coordinates is convergent in \mathbb{C} . The notation gets a little messy: if $x_n = (a_{n1}, a_{n2}, \dots, a_{nr})$ and $x = (a_1, \dots, a_r)$, so that

$$d_\infty(x_n, x) = \max_{1 \leq k \leq r} |a_{nk} - a_k|,$$

then $x_n \rightarrow x$ for d_∞ if and only if, for each $k = 1, \dots, r$, $a_{nk} \rightarrow a_k$ in \mathbb{C} .

(v) If $p \geq 1$, d_p is the Minkowski p -metric on \mathbb{C}^r (3.1.14) and d_∞ is the sup-metric of the preceding example, then

$$d_\infty(x, y) \leq d_p(x, y) \leq r^{1/p} d_\infty(x, y)$$

for all x, y in \mathbb{C}^r (briefly, $d_\infty \leq d_p \leq r^{1/p} d_\infty$).

{Proof: If $x = (a_1, \dots, a_r)$, $y = (b_1, \dots, b_r)$ and j is an index such that $|a_j - b_j| = \|x - y\|_\infty$, then

$$(\|x - y\|_\infty)^p = |a_j - b_j|^p \leq \sum_{k=1}^r |a_k - b_k|^p \leq r \left(\max_{1 \leq k \leq r} |a_k - b_k| \right)^p,$$

therefore $\|x - y\|_\infty \leq \|x - y\|_p \leq r^{1/p} \|x - y\|_\infty$.

It follows that

$$x_n \rightarrow x \text{ for } d_\infty \Leftrightarrow x_n \rightarrow x \text{ for } d_p;$$

in view of Example (iv), convergence for d_p means coordinatewise convergence.

These assertions hold a fortiori with \mathbb{C}^r replaced by \mathbb{R}^r ; in particular, in $\mathbb{C} = \mathbb{R}^2$, a sequence is convergent for d_p (or d_∞) if and only if its real and imaginary parts are convergent in \mathbb{R} .

(vi) Let T be a nonempty set, d_∞ the sup-metric on $\mathcal{B} = \mathcal{B}_\mathbb{C}(T)$. If $x_n, x \in \mathcal{B}$ then

$$x_n \rightarrow x \text{ for } d_\infty \Leftrightarrow (\forall \epsilon > 0) \sup_{t \in T} |x_n(t) - x(t)| \leq \epsilon \text{ ultimately}$$

$$\Leftrightarrow (\forall \epsilon > 0) \exists N \ni n \geq N \Rightarrow |x_n(t) - x(t)| \leq \epsilon \text{ for all } t \in T;$$

one then says that $x_n \rightarrow x$ *uniformly* on T . Since

$$|x_n(t) - x(t)| \leq d_\infty(x_n, x)$$

for each $t \in T$, it follows that

$$x_n \rightarrow x \text{ uniformly} \Rightarrow (\forall t \in T) \ x_n(t) \rightarrow x(t) \text{ in } \mathbb{C};$$

so to speak, uniform convergence implies 'pointwise convergence'. {The converse is false (Exercise 1).}

3.2.3. Proposition. If d is a pseudometric on a set X , then

$$|d(x, y) - d(x', y')| \leq d(x, x') + d(y, y')$$

for all x, y, x', y' in X .

Proof. By the triangle inequality, $d(x, y) \leq d(x, x') + d(x', y') + d(y', y)$, thus

$$d(x, y) - d(x', y') \leq d(x, x') + d(y, y');$$

interchanging $x \leftrightarrow x'$ and $y \leftrightarrow y'$ yields

$$d(x', y') - d(x, y) \leq d(x, x') + d(y, y'),$$

whence the asserted inequality. \diamond

3.2.4. Corollary. In a metric space (X, d) , if $x_n \rightarrow x$ and $y_n \rightarrow y$ in X , then $d(x_n, y_n) \rightarrow d(x, y)$ in \mathbb{R} .

Proof. $|d(x_n, y_n) - d(x, y)| \leq d(x_n, x) + d(y_n, y)$. \diamond

3.2.5. *Definition.* Let (X, d) be a metric space. A subset A of X is said to be **closed** in X (or to be a *closed set*) if, whenever the terms of a convergent sequence are in A , then the limit must also be in A ; in symbols,

$$\left. \begin{array}{l} a_n \in A \\ x \in X \\ a_n \rightarrow x \end{array} \right\} \Rightarrow x \in A.$$

{Mnemonic: "You can't fight your way out of a closed set with a convergent sequence."}

3.2.6. *Example.* (i) In \mathbb{R} with the usual metric, every closed interval $[a, b]$ is a closed set. {If $a \leq x_n \leq b$ and $x_n \rightarrow x$ then $a \leq x \leq b$ (1.8.24).}

(ii) In a discrete metric space, every subset is a closed set (cf. 3.2.2, (ii)).

(iii) If (X, d) is a metric space and $D = d/(1 + d)$, then a subset A of X is closed for d if and only if it is closed for D (cf. 3.2.2, (iii)).

(iv) In any metric space, singletons $A = \{a\}$ are closed sets.

(v) If A_1, \dots, A_r are closed subsets of \mathbb{C} (for the usual metric) then the product set $A_1 \times \dots \times A_r$ is a closed subset of \mathbb{C}^r for the metrics d_∞ and d_p (cf. 3.2.2, (iv), (v)).

(vi) Let (X, d) be a metric space. For every $c \in X$ and every real number $r > 0$, the set

$$B_r(c) = \{x \in X : d(x, c) \leq r\}$$

is a closed set. For, if $x_n \rightarrow x$ and $d(x_n, c) \leq r$ for all n , then $d(x, c) \leq r$ by 3.2.4 and part (i).

For example, if $\mathcal{B} = \mathcal{B}_{\mathbb{C}}(T)$ as in 3.2.2, (vi), then, for every $r > 0$, the set

$$\{x \in \mathcal{B} : |x(t)| \leq r \text{ for all } t \in T\}$$

is closed for the sup-metric d_∞ .

3.2.7. *Definition.* With notations as in (vi) above, the set $B_r(c)$ is called the **closed ball** in X with **center** c and **radius** r .

As noted in 3.2.6, (vi), every closed ball in a metric space is a closed set. In \mathbb{R}^2 with the Euclidean metric (3.1.15), 'closed balls' in the sense of 3.2.7 are discs; in Euclidean 3-space, this use of the word 'ball' accords with its meaning in everyday language.

The key formal properties of closed sets are as follows:

3.2.8. *Theorem.* In a metric space (X, d) ,

- (1) \emptyset and X are closed sets;
- (2) the intersection of any family of closed sets is a closed set;
- (3) the union of any two closed sets is a closed set.

Proof. (1) The empty set is closed ‘by default’ (no sequence exists that contradicts the assertion that \emptyset is closed), and it is trivial that X is closed.

(2) Let $A = \bigcap_{i \in I} A_i$, where $(A_i)_{i \in I}$ is a family of closed sets. If (x_n) is a sequence in A and $x_n \rightarrow x$ then, for every index $i \in I$, $x \in A_i$ (because $x_n \in A_i$ for all n , and A_i is closed), therefore $x \in A$.

(3) Let A and B be closed sets and suppose $x_n \rightarrow x$, where $x_n \in A \cup B$ for all n . If $x_n \in A$ frequently, then $x_{n_k} \in A$ for a subsequence (x_{n_k}) , whence $x \in A$ (because $x_{n_k} \rightarrow x$ and A is closed); the alternative is that $x_n \in B$ ultimately, which implies that $x \in B$. Either way, $x \in A \cup B$, thus $A \cup B$ is closed. \diamond

3.2.9. Corollary. *If A_1, \dots, A_n are closed subsets of X then so is $A_1 \cup \dots \cup A_n$. In particular, every finite subset of X is closed.*

Proof. Induction on (3), then cite 3.2.6, (iv). \diamond

In \mathbb{R} with the usual metric, if $a < b$ then the interval $[a, b]$ is, according to 3.2.7, a ‘closed ball’ with center $c = \frac{1}{2}(a + b)$ and radius $r = \frac{1}{2}(b - a)$:

$$[a, b] = \{x \in \mathbb{R} : |x - c| \leq r\} = B_r(c).$$

For the open interval, the inequality is strict:

$$(a, b) = \{x \in \mathbb{R} : |x - c| < r\};$$

by analogy, one defines ‘open balls’ in any metric space:

3.2.10. Definition. Let (X, d) be a metric space, $c \in X$, $r > 0$. The set

$$U_r(c) = \{x \in X : d(x, c) < r\}$$

is called the **open ball** in X with center c and radius r . Note that $c \in U_r(c) \subset B_r(c)$.

3.2.11. Examples. (i) In a discrete metric space (X, d) , if $0 < r \leq 1$ then $U_r(c) = \{c\}$ for all $c \in X$; if $r > 1$ then $U_r(c) = X$ for every c .

(ii) In any metric space (X, d) , every closed ball is the intersection of a sequence of open balls:

$$B_r(c) = \bigcap_{n=1}^{\infty} U_{r+1/n}(c),$$

because $d(x, c) \leq r \Leftrightarrow d(x, c) < r + 1/n$ for all n . {Exercise: Interpret this formula for the case that d is the discrete metric on X and, in turn, $0 < r < 1$, $r = 1$, $r > 1$.}

3.2.12. Definition. Let (X, d) be a metric space, A a subset of X , c a point of A . We say that c is **interior** to A (or is an *interior point* of A) if A contains an open ball with center c , that is, if there exists

an $r > 0$ such that $U_r(c) \subset A$. (So to speak, not only is c in A , but a 'buffer zone' around c is contained in A ; but this is just talk, and in a discrete metric space it is nonsense.) If c is interior to A , we also say that A is a **neighborhood** of c .

We say that A is an **open set** in X if every point of A is an interior point:

$$(\forall c \in A) \exists r > 0 \ni U_r(c) \subset A$$

(note that the size of r may vary with c); thus, A is open if and only if it is a neighborhood of each of its points.

3.2.13. *Remark.* With the preceding notations,

- (i) \emptyset and X are open sets (by default and trivially, respectively);
- (ii) if V_1 and V_2 are neighborhoods of c , then so is $V_1 \cap V_2$ (think $\min\{r_1, r_2\}$);
- (iii) if V is a neighborhood of c and if $W \supset V$, then W is also a neighborhood of c ;
- (iv) for every $r > 0$, $U_r(c)$ is a neighborhood of c . Better yet:

3.2.14. *Proposition.* In a metric space (X, d) , every open ball $U_r(c)$ is an open set.

Proof. Let $x \in U_r(c)$; we are to show that x is interior to $U_r(c)$, thus we seek an $s > 0$ such that $U_s(x) \subset U_r(c)$. Let $s = r - d(x, c)$, which is > 0 because $x \in U_r(c)$. If $y \in U_s(x)$ then

$$d(y, c) \leq d(y, x) + d(x, c) < s + d(x, c) = r,$$

therefore $y \in U_r(c)$. (Draw a picture!) \diamond

Here are the key properties of open sets:

3.2.15. *Theorem.* In a metric space (X, d) ,

- (1) \emptyset and X are open sets;
- (2) the union of any family of open sets is open;
- (3) the intersection of any two open sets is open.

Proof. (1) Noted in 3.2.13.

(2) Let $(U_i)_{i \in I}$ be a family of open sets and let $U = \bigcup_{i \in I} U_i$. If $x \in U$ then $x \in U_i$ for some i , therefore U_i is a neighborhood of x , hence so is its superset U ; thus U is a neighborhood of each of its points.

(3) Let U and V be open sets. If $x \in U \cap V$ then $x \in U$ and $x \in V$, so U and V are neighborhoods of x , hence so is $U \cap V$. \diamond

The preceding theorem (properties of open sets) is hauntingly similar to Theorem 3.2.8 (properties of closed sets): to get the open set theorem from

the closed set theorem, we need only make the following substitutions:

$$\begin{aligned} \text{intersection} &\mapsto \text{union} \\ \text{union} &\mapsto \text{intersection} \\ \emptyset &\mapsto X \\ X &\mapsto \emptyset \\ \text{closed} &\mapsto \text{open}. \end{aligned}$$

The first two of these transformations are accomplished by complementation (De Morgan's formulas), as are the next two; so is the last:

3.2.16. Theorem. *For a subset A of a metric space,*

$$A \text{ is closed} \Leftrightarrow \mathbb{C}A \text{ is open}.$$

Proof. The arguments go more smoothly by proving the contrapositive form:

$$A \text{ is not closed} \Leftrightarrow \mathbb{C}A \text{ is not open}.$$

\Rightarrow : Assuming A is not closed, there exist a sequence (a_n) in A and a point $x \in \mathbb{C}A$ such that $a_n \rightarrow x$. It will suffice to show that x is not an interior point of $\mathbb{C}A$. Indeed, for every $r > 0$, the existence of indices n such that $d(a_n, x) < r$ shows that $\mathbb{C}A$ fails to contain the open ball $U_r(x)$.

\Leftarrow : Assuming $\mathbb{C}A$ is not open, let x be a point of $\mathbb{C}A$ that is not an interior point. For every positive integer n , $\mathbb{C}A$ fails to contain the open ball $U_{1/n}(x)$, so there exists a point $a_n \in U_{1/n}(x)$ with $a_n \notin \mathbb{C}A$; then $a_n \in A$ and $a_n \rightarrow x \notin A$, which shows that A is not closed. \diamond

3.2.17. Corollary. $A \text{ is open} \Leftrightarrow \mathbb{C}A \text{ is closed}.$

Proof. By the theorem, $\mathbb{C}A \text{ is closed} \Leftrightarrow \mathbb{C}(\mathbb{C}A) \text{ is open}.$ \diamond

3.2.18. Corollary. *For every $c \in X$ and $r > 0$, the set $S_r(c) = \{x \in X : d(x, c) = r\}$ is closed.*

Proof. $S_r(c) = \{x : d(x, c) \leq r\} - \{x : d(x, c) < r\} = B_r(c) \cap (\mathbb{C}U_r(c))$ is the intersection of two closed sets (3.2.6, (vi) and 3.2.17). \diamond

The set $S_r(c)$ of 3.2.18 is called the **sphere** with center c and radius r . {In \mathbb{R} with the usual metric, a 'sphere' in this sense is a pair of points; in the Euclidean plane, it is a circle; in Euclidean 3-space, it is an honest sphere—the surface of an honest ball.}

CAUTION: Spheres can be empty and the points of $S_r(c)$ need not be limits of points of the open ball $U_r(c)$ (think discretely!).

Sequential convergence in a metric space is neatly described in terms of open sets:

3.2.19. Proposition. *Let (X, d) be a metric space, (x_n) a sequence of points in X , and $x \in X$. The following conditions are equivalent:*

- (a) $x_n \rightarrow x$ for d ;
 (b) if U is an open set containing x , then $x_n \in U$ ultimately.

Proof. (a) \Rightarrow (b): Choose $r > 0$ so that $U_r(x) \subset U$ and note that $d(x_n, x) < r$ ultimately.

(b) \Rightarrow (a): For every $\epsilon > 0$ consider $U = U_\epsilon(x)$. \diamond

For a reverse spin on this proposition, see Exercise 4.

3.2.20. We mention, for future use (5.3.2, 6.1.14, 6.3.1), another metric concept with a geometric flavor: if A is a nonempty subset of a metric space (X, d) , the **diameter** of A , denoted $\text{diam } A$, is defined by

$$\text{diam } A = \sup\{d(x, y) : x, y \in A\}.$$

In general, $0 \leq \text{diam } A \leq +\infty$; if $\text{diam } A < +\infty$, A is said to have *finite diameter*.

Exercises

1. For each positive integer n , let $x_n : [0, 1] \rightarrow \mathbb{R}$ be the piecewise linear function such that $x_n(0) = 0$, $x_n(1/2n) = 1$ and $x_n = 0$ on $[1/n, 1]$ (draw a picture). Then $x_n \rightarrow 0$ pointwise on $[0, 1]$, but $\|x_n\|_\infty = 1$ for all n .

2. Give a proof of the open set properties (3.2.15) based on the characterization of open sets in 3.2.16.

3. In a metric space, every closed set is the intersection of a sequence of open sets.

4. Let (X, d) be a metric space, U a subset of X . The following conditions on U are equivalent: (a) U is open; (b) if $x_n \rightarrow x \in U$ then $x_n \in U$ ultimately.

5. Let (X, d) be a metric space, A a subset of X . The following conditions on A are equivalent: (a) A is closed; (b) if $x_n \rightarrow x$ and $x_n \in A$ frequently, then $x \in A$.

3.3. Topological Spaces

The properties (1)–(3) in the ‘open set theorem’ (3.2.15) are stated without overt reference to the metric in question; this is possible because there is a prior understanding as to the meaning of ‘open set’ (which *does* entail the metric). Many useful metric space concepts (such as limit and continuity), as well as their analogues in situations where there is no metric, can be formulated in terms of these three properties. Efficiency suggests abstracting

the notion of open set (i.e., taking it to be an undefined term) and taking the statements (1)–(3) to be the axioms for an abstract structure, the consequences of the axioms then being applicable to all concrete situations where the axioms are verified. Such structures are called topological spaces:

3.3.1. *Definition.* Let X be a set, \mathcal{O} a set of subsets of X (thus $\mathcal{O} \subset \mathcal{P}(X)$, where $\mathcal{P}(X)$ is the power set of X). One says that \mathcal{O} is (or ‘defines’) a **topology** on X if it satisfies the following three conditions:

- 1° $\emptyset \in \mathcal{O}$ and $X \in \mathcal{O}$;
- 2° if $(U_i)_{i \in I}$ is a family of sets in \mathcal{O} , then $\bigcup_{i \in I} U_i \in \mathcal{O}$;
- 3° if $U, V \in \mathcal{O}$ then $U \cap V \in \mathcal{O}$.

The sets in \mathcal{O} are called the **open sets** of X (for the topology \mathcal{O}) and the pair (X, \mathcal{O}) is called a **topological space**; one also speaks of X as a topological space if it is not necessary to have an explicit notation for its set of open sets.

In words, the conditions 1°–3° say that the empty set and X itself are open sets, the union of any family of open sets is open, and the intersection of any two (hence any finite number of) open sets is open.

Guided by 3.2.16, a set $A \subset X$ is said to be **closed** if its complement is open; that is (by definition), A closed $\Leftrightarrow \complement A \in \mathcal{O}$. {In the context of metric spaces, the terms ‘closed’ and ‘open’ were defined independently, consequently the statement ‘ A closed $\Leftrightarrow \complement A$ open’ was then a theorem (3.2.16) rather than a definition.}

3.3.2. *Example.* (i) If d is a metric on a set X , we write \mathcal{O}_d for the topology on X derived from d as in 3.2.15; the sets $U \in \mathcal{O}_d$ are said to be open for d (or d -open). In general, a topological space (X, \mathcal{O}) is said to be *metrizable* if there exists a metric d on X such that $\mathcal{O}_d = \mathcal{O}$. Different metrics may produce the same topology (for instance d and $2d$), and not every topological space is metrizable (Exercise 1).

(ii) For every set X , the set $\mathcal{O} = \{\emptyset, X\}$ is a topology on X , called—deservedly—the *trivial* topology on X .

(iii) For every set X , the set $\mathcal{O} = \mathcal{P}(X)$ is a topology on X , called the *discrete* topology on X . {Exercise: Justify the name (cf. 3.1.7) by showing that a discrete topological space is metrizable.}

3.3.3. *Definition.* Two metrics d, d' on a set X are said to be **equivalent** (or ‘topologically equivalent’) if they define the same topology, that is, $\mathcal{O}_d = \mathcal{O}_{d'}$.

On the way to characterizing equivalence in terms of convergent sequences:

3.3.4. *Lemma.* Let (X, d) be a metric space, A a subset of X , and x a point of A . The following conditions are equivalent:

- (a) x is an interior point of A ;
- (b) $d(x_n, x) \rightarrow 0 \Rightarrow x_n \in A$ ultimately.

Proof. (a) \Rightarrow (b): This is immediate from 3.2.19.

$\sim(a) \Rightarrow \sim(b)$: Suppose the point x of A is not an interior point. For every positive integer n , A fails to contain the open ball $U_{1/n}(x)$, so there exists a point $x_n \in \complement A$ such that $d(x_n, x) < 1/n$. Then $d(x_n, x) \rightarrow 0$ but x_n does not belong to A for any n , let alone ultimately, whence $\sim(b)$. \diamond

3.3.5. Proposition. For a pair of metrics d, d' on a set X , the following conditions are equivalent:

- (a) $d(x_n, x) \rightarrow 0 \Rightarrow d'(x_n, x) \rightarrow 0$.
- (b) $\mathcal{O}_d \supset \mathcal{O}_{d'}$.

Proof. (a) \Rightarrow (b): Let U' be open for the d' -topology; assuming (a), we are to show that U' is d -open. Let $x \in U'$ and suppose $d(x_n, x) \rightarrow 0$. By (a), $d'(x_n, x) \rightarrow 0$, and since U' is d' -open we have $x_n \in U'$ ultimately, by the lemma; again citing the lemma (this time in the other direction) we infer that x is interior to U' for the d -topology. Thus every point of U' is an interior point for the topology \mathcal{O}_d , therefore $U' \in \mathcal{O}_d$ (3.2.12).

(b) \Rightarrow (a): Let $d(x_n, x) \rightarrow 0$; assuming (b), we are to show that $d'(x_n, x) \rightarrow 0$. Given any $\epsilon > 0$, we must show that $d'(x_n, x) < \epsilon$ ultimately. Let $U = \{y \in X : d'(y, x) < \epsilon\}$. Since $U \in \mathcal{O}_{d'}$ (3.2.14) we have $U \in \mathcal{O}_d$ by (b), so it follows from $d(x_n, x) \rightarrow 0$ and 3.2.19 that $x_n \in U$ ultimately, as we wished to show. \diamond

It follows at once that a metric topology is determined by the associated concept of sequential convergence:

3.3.6. Corollary. For a pair of metrics d, d' on a set X , the following conditions imply each other:

- (a) d and d' are equivalent (that is, $\mathcal{O}_d = \mathcal{O}_{d'}$);
- (b) $d(x_n, x) \rightarrow 0 \Leftrightarrow d'(x_n, x) \rightarrow 0$.

Proof. Condition (a) says that $\mathcal{O}_d \supset \mathcal{O}_{d'}$ and $\mathcal{O}_{d'} \supset \mathcal{O}_d$. \diamond

3.3.7. Corollary. (i) For any metric space (X, d) , the metrics d and $D = d/(1 + d)$ are equivalent; thus every metric is equivalent to a metric with values in $[0, 1)$.

(ii) On \mathbb{C}^r , the metrics d_∞ and d_p ($p \geq 1$) of 3.1.10, (ii) and 3.1.14 are all equivalent.

Proof. Immediate from 3.2.2, (iii), (v) and the preceding corollary. \diamond

3.3.8. Definition. Let (X, \mathcal{O}) be a topological space, $c \in A \subset X$. Guided by the case of metric spaces (3.2.12) we say that x is **interior** to A (or that A is a **neighborhood** of x) if there exists an open set U such that $c \in U \subset A$ (so to speak, an open set can be interpolated between c and A).

3.3.9. *Remarks.* In a topological space (X, \mathcal{O}) :

(i) A set U is open (that is, $U \in \mathcal{O}$) if and only if U is a neighborhood of each of its points. {“If”: For each $x \in U$ choose an open set U_x such that $x \in U_x \subset U$; then $U = \bigcup_{x \in U} U_x$ is the union of a family of open sets.}

(ii) If V_1 and V_2 are neighborhoods of x , then so is $V_1 \cap V_2$. {Reason: The intersection of two open sets containing x is an open set containing x .}

(iii) If $V \subset W \subset X$ and V is a neighborhood of x , then so is W .

3.3.10. *Example.* Let (X, d) be a metric space and let $x \in X$. For every positive integer n , let $V_n = U_{1/n}(x)$. Then the V_n are neighborhoods of x , and every neighborhood V of x contains V_n for some n . {For, if $U_r(x) \subset V$ and $1/n < r$ then $V_n \subset U_r(x) \subset V$.}

3.3.11. *Definition.* In a topological space X , a **fundamental sequence** of neighborhoods of a point x is a sequence (V_n) of neighborhoods of x such that every neighborhood of x contains (at least) one of the V_n . If every point of X has a fundamental sequence of neighborhoods, then X is said to be **first countable**.

Every metric space is first countable (3.3.10). There exist topological spaces in which no point has a fundamental sequence of neighborhoods (Exercise 3).

3.3.12. *Definition.* Let A be a subset of a topological space X . A point x of X is said to be **adherent** to A (or to be an ‘adherent point’ of A) if every neighborhood of x intersects A , that is, if

$$V \text{ a neighborhood of } x \Rightarrow V \cap A \neq \emptyset.$$

(It is the same to say that every open set containing x intersects A .)

3.3.13. *Examples.* (i) In \mathbb{R} with the usual (metric) topology, if I is an interval with endpoints $a \leq b$, the points adherent to I are precisely the points of the closed interval $[a, b]$. {Proof: If, for example, $x > b$, and if $r = x - b$, then $(x - r, x + r)$ is a neighborhood of x disjoint from I , therefore x is not adherent to I .} Every $x \in \mathbb{R}$ is adherent to the set \mathbb{Q} of rational numbers (because every nondegenerate interval contains a rational number).

(ii) In a metric space (X, d) , x is adherent to A if and only if there exists a sequence (a_n) in A such that $a_n \rightarrow x$. {“Only if”: Assuming x adherent to A , for every positive integer n the neighborhood $U_{1/n}(x)$ of x contains a point a_n of A .}

(iii) In a discrete topological space (3.3.2, (iii)), x is adherent to A if and only if $x \in A$. {For, if $x \notin A$ then $\{x\}$ is a neighborhood of x disjoint from A .}

(iv) In a trivial topological space (3.3.2, (ii)), every $x \in X$ is adherent to every nonempty subset A of x . {Reason: X is the only available neighborhood of x .}

3.3.14. *Definition.* If A is a subset of a topological space X , the set of all points x of X that are adherent to A is called the **closure** (or **adherence**) of A and is denoted \overline{A} :

$$\overline{A} = \{x \in X : x \text{ is adherent to } A\}.$$

(The reason for the term 'closure' is clear from (5) of 3.3.16 below.)

3.3.15. *Examples.* (i) In \mathbb{R} with the usual topology, $\overline{\mathbb{Q}} = \mathbb{R}$, and if I is an interval with endpoints $a \leq b$, then $\overline{I} = [a, b]$ (cf. 3.3.13, (i)).

(ii) For a subset A of a metric space (X, d) ,

$$\overline{A} = \{x \in X : a_n \rightarrow x \text{ for some sequence } (a_n) \text{ in } A\}$$

(cf. 3.3.13, (ii)).

(iii) In a discrete topological space, $\overline{A} = A$ for every subset A (3.3.13, (iii)).

(iv) In a trivial topological space, $\overline{A} = X$ for every nonempty subset A (3.3.13, (iv)).

In a topological space X , $A \mapsto \overline{A}$ is a mapping $\mathcal{P}(X) \rightarrow \mathcal{P}(X)$; here are its key properties:

3.3.16. *Theorem.* Let X be a topological space, A and B subsets of X . Then:

- (1) $A \subset \overline{A}$;
- (2) $A \subset B \Rightarrow \overline{A} \subset \overline{B}$;
- (3) $\overline{\overline{A}} = \overline{A}$;
- (4) $A = \overline{A} \Leftrightarrow A$ is a closed set;
- (5) \overline{A} is the smallest closed set containing A ;
- (6) $\overline{A \cup B} = \overline{A} \cup \overline{B}$.

Proof. (1) If $a \in A$ and V is a neighborhood of a , then $V \cap A \neq \emptyset$ (because $a \in V$).

(2) If every neighborhood of x intersects A then it intersects B .

(3) By $\overline{\overline{A}}$ we mean the closure of \overline{A} , that is, $\overline{\overline{A}} = (\overline{A})^-$. From $A \subset \overline{A}$ we have $\overline{A} \subset \overline{\overline{A}}$ by (2). To prove the reverse inclusion, assuming $x \in \overline{\overline{A}}$ and V a neighborhood of x , we must show that $V \cap A \neq \emptyset$. We can suppose that V is open (3.3.8). Since x is adherent to \overline{A} , $V \cap \overline{A} \neq \emptyset$; say $y \in V \cap \overline{A}$. Then $y \in \overline{A}$ and V is also a neighborhood of y , therefore $V \cap A \neq \emptyset$.

(4) In any case $A \subset \overline{A}$, so the problem is to show that

$$A \supset \overline{A} \Leftrightarrow A \text{ is closed,}$$

in other words,

$$\mathcal{C}A \subset \overline{\mathcal{C}A} \Leftrightarrow \mathcal{C}A \text{ is open.}$$

To say that $\mathcal{C}A \subset \mathcal{C}\bar{A}$ means that no point of $\mathcal{C}A$ is adherent to A ; by the definition of adherent point, this means that

$$x \in \mathcal{C}A \Rightarrow \text{some neighborhood of } x \text{ is disjoint from } A,$$

in other words,

$$x \in \mathcal{C}A \Rightarrow \text{some neighborhood of } x \text{ is contained in } \mathcal{C}A,$$

and it is the same to say that

$$x \in \mathcal{C}A \Rightarrow \mathcal{C}A \text{ is a neighborhood of } x.$$

The latter implication says that $\mathcal{C}A$ is an open set (3.3.9, (i)).

(5) Since $\bar{\bar{A}} = \bar{A} \supset A$, \bar{A} is a closed set containing A . On the other hand, if B is a closed set containing A then $\bar{A} \subset \bar{B} = B$, whence the minimality of \bar{A} .

(6) Since $A \subset \bar{A}$ and $B \subset \bar{B}$ we have $A \cup B \subset \bar{A} \cup \bar{B}$, and since $\bar{A} \cup \bar{B}$ is closed (it is the union of two closed sets) it follows from (5) that $\overline{A \cup B} \subset \bar{A} \cup \bar{B}$. On the other hand, $A \subset A \cup B$ and $B \subset A \cup B$ imply $\bar{A} \subset \overline{A \cup B}$ and $\bar{B} \subset \overline{A \cup B}$, therefore $\bar{A} \cup \bar{B} \subset \overline{A \cup B}$. \diamond

3.3.17. *Example.* (The metric topology of $\bar{\mathbb{R}}$) In §1.15 we introduced the notation $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$; we are going to show that there is a metric topology on $\bar{\mathbb{R}}$ for which, among other things, $\bar{\mathbb{R}}$ is the closure of its subset \mathbb{R} (thus justifying the notation of §1.15). The first step is to note that $\bar{\mathbb{R}}$ is order-isomorphic to a closed interval. As in 1.16.10, define a function $f: [-1, 1] \rightarrow \bar{\mathbb{R}}$ by the formulas

$$f(x) = \begin{cases} -\infty & \text{for } x = -1 \\ \frac{x}{1-x^2} & \text{for } -1 < x < 1 \\ +\infty & \text{for } x = 1. \end{cases}$$

For every $x \in (-1, 1)$,

$$f'(x) = \frac{1+x^2}{(1-x^2)^2} > 0,$$

so f is strictly increasing on $(-1, 1)$; ¹ since f is unbounded in the neighborhood of ± 1 , it follows that f is a strictly increasing bijection $[-1, 1] \rightarrow \bar{\mathbb{R}}$. The formula

$$D(y, z) = |f^{-1}(y) - f^{-1}(z)|$$

defines a metric on $\bar{\mathbb{R}}$. We now have two notions of sequential convergence in $\bar{\mathbb{R}}$: the order-theoretic convergence $y_n \rightarrow y$ defined in §1.16, and convergence $y_n \rightarrow y$ for the metric D ; we are going to show that they are the same.

¹First course, p. 130, Theorem 8.5.2.

Since f is an order-isomorphism $[-1, 1] \rightarrow \overline{\mathbb{R}}$, for every sequence (x_n) in $[-1, 1]$ we have

$$\begin{aligned} f(\limsup x_n) &= \limsup f(x_n), \\ f(\liminf x_n) &= \liminf f(x_n), \end{aligned}$$

therefore

$$\liminf x_n = \limsup x_n \Leftrightarrow \liminf f(x_n) = \limsup f(x_n),$$

in other words,

$$(x_n) \text{ is convergent in } [-1, 1] \Leftrightarrow (f(x_n)) \text{ is convergent in } \overline{\mathbb{R}} \\ \text{(as in §1.16).}$$

Thus, for $x_n, x \in [-1, 1]$,

$$(*) \quad x_n \rightarrow x \text{ in } [-1, 1] \Leftrightarrow f(x_n) \rightarrow f(x) \text{ in } \overline{\mathbb{R}},$$

where on the left side we have convergence for the usual metric, and on the right side convergence in the sense of §1.16.

Suppose now that $y_n, y \in \overline{\mathbb{R}}$. Then

$$\begin{aligned} D(y_n, y) \rightarrow 0 &\Leftrightarrow |f^{-1}(y_n) - f^{-1}(y)| \rightarrow 0 \\ &\Leftrightarrow f^{-1}(y_n) \rightarrow f^{-1}(y) \text{ in } [-1, 1] \\ &\Leftrightarrow f(f^{-1}(y_n)) \rightarrow f(f^{-1}(y)) \text{ in } \overline{\mathbb{R}} \quad [\text{by } (*)] \\ &\Leftrightarrow y_n \rightarrow y \text{ in } \overline{\mathbb{R}} \quad [\text{in the sense of §1.16}]; \end{aligned}$$

thus sequential convergence for the metric D coincides with the concept of convergence defined in §1.16.

In particular, if $y_n, y \in \mathbb{R}$ then $y_n \rightarrow y$ for the usual metric if and only if $y_n \rightarrow y$ in the sense of §1.16 (cf. 1.16.8), in other words $D(y_n, y) \rightarrow 0$; this shows that the restriction of D to $\mathbb{R} \times \mathbb{R}$ is equivalent to the usual metric on \mathbb{R} (3.3.6).

Finally, since $n \rightarrow +\infty$ and $-n \rightarrow -\infty$ (for D or in the sense of §1.16—it is the same), it follows that the closure of \mathbb{R} for the D -topology on $\overline{\mathbb{R}}$ is equal to $\overline{\mathbb{R}}$. Thus \mathbb{R} is dense in $\overline{\mathbb{R}}$ in the sense of the following definition:

3.3.18. Definition. A set A in a topological space X is said to be **dense** (in X) if $\overline{A} = X$.

Exercises

1. If a set X has two or more points, then the trivial topology on X is not metrizable.

{Hint: 3.2.6, (iv).}

2. Two metrics d and d' on a set X are equivalent if and only if they satisfy the following condition: (c) a sequence in X is convergent for $d \Leftrightarrow$ it is convergent for d' .

{Note the subtle difference between (c) and the condition (b) of 3.3.6: we must contemplate—and rule out—the possibility that the d -limit might be different from the d' -limit.}

3. Let \mathcal{O} be the set of all subsets A of \mathbb{R} such that $\mathbb{R} - A$ is finite, together with the empty set.

(i) \mathcal{O} is a topology on \mathbb{R} .

(ii) The topological space $(\mathbb{R}, \mathcal{O})$ is not first countable (in the sense of 3.3.11).

{Hint: (ii) If (V_n) were a fundamental sequence of neighborhoods of x , one would have $\bigcap_{n=1}^{\infty} V_n = \{x\}$. Now take complements.}

4. A sequence (x_n) in a topological space is said to be *convergent* (for the given topology) if there exists a point x in the space such that, for every neighborhood V of x , $x_n \in V$ ultimately; if \mathcal{O} is the topology in question, we then say that $x_n \rightarrow x$ for \mathcal{O} (x is not asserted to be unique). For a metric space, this coincides with the usual notion of sequential convergence (3.2.19).

Let (X, d) be a metric space and let \mathcal{O} be a topology on X . The following conditions are equivalent:

(a) $\mathcal{O} = \mathcal{O}_d$;

(b) the topological space (X, \mathcal{O}) is first countable, and $x_n \rightarrow x$ for $\mathcal{O} \Leftrightarrow d(x_n, x) \rightarrow 0$.

5. Let X be a topological space, A a subset of X . The set of all interior points of A (in the sense of 3.3.8) is called the **interior** of A , written A° (or $\text{int } A$); thus,

$$A^\circ = \{x \in X : x \text{ is interior to } A\}.$$

Prove that $x \notin \overline{A} \Leftrightarrow x \in (\mathbb{C}A)^\circ$, in other words $\overline{\mathbb{C}A} = (\mathbb{C}A)^\circ$, and infer that $A^\circ = \mathbb{C}(\overline{\mathbb{C}A})$.

6. Let X be a topological space, A a subset of X . A point $x \in X$ is said to be **exterior** to A if it is interior to $\mathbb{C}A$, in other words, if there exists a neighborhood of x that is disjoint from A ; the set of all such x is called the **exterior** of A , written $\text{ext } A$, thus $\text{ext } A = \text{int}(\mathbb{C}A)$. A point of X that is neither interior nor exterior to A is called a **boundary point** of A , and the set of all such points is called the **boundary** of A , written ∂A (or $\text{bd } A$); thus

$$X = \text{int } A \cup \text{ext } A \cup \text{bd } A$$

is a partition of X into three pairwise disjoint subsets.

(i) Show that $\partial A = \overline{A} \cap \overline{\mathbb{C}A} = \overline{A} - A^\circ = \partial(\mathbb{C}A)$.

(ii) If $X = \mathbb{R}$ and $A = \mathbb{Q}$, describe ∂A . What if $A = \mathbb{Z}$? What if $A = \mathbb{R}$?

(iii) Show that if A is a nonempty subset of \mathbb{R} distinct from \mathbb{R} , then A has nonempty boundary.

7. Let (X, \mathcal{O}) be a topological space, let A be a subset of X and let \mathcal{O}_A be the set of all intersections of A with the sets of \mathcal{O} :

$$\mathcal{O}_A = \mathcal{O} \cap A = \{U \cap A : U \in \mathcal{O}\}.$$

(i) \mathcal{O}_A is a topology on A . (\mathcal{O}_A is called the *relative topology* on A induced by the topology of X ; when A is equipped with the relative topology, it is called a *subspace* of the topological space X .)

(ii) If \mathcal{C} is the set of all closed sets in X then the closed sets in the topological space (A, \mathcal{O}_A) are the sets $C \cap A$ ($C \in \mathcal{C}$).

(iii) Let $B \subset A$. We write $\text{cl}_A B$ and $\text{int}_A B$ for the closure and interior of B for the relative topology on A . Then

$$\text{cl}_A B = \overline{B} \cap A \quad \text{and} \quad \text{int}_A B \supset B^\circ \cap A$$

(where \overline{B} and B° are the closure and interior of B in X); the inclusion may be proper.

8. Let (X, d) be a metric space and let A be a nonempty subset of X . The restriction of the function $d : X \times X \rightarrow [0, +\infty)$ to $A \times A$ is a metric on A ; let us denote it by d_A . The metric space (A, d_A) is called a *metric subspace* of X . The topology on A derived from d_A coincides with the relative topology $\mathcal{O}_d \cap A$ induced by \mathcal{O}_d (Exercise 7).

3.4. Continuity

3.4.1. *Definition.* Let X and Y be metric spaces, let $f : X \rightarrow Y$ and let c be a point of X . The function f is said to be **continuous at c** if

$$x_n \rightarrow c \Rightarrow f(x_n) \rightarrow f(c).$$

3.4.2. *Proposition.* With notations as in 3.4.1, let d be the given metric on X , ρ the given metric on Y . The following conditions are equivalent:

- (a) f is continuous at c ;
- (b) W a neighborhood of $f(c) \Rightarrow f^{-1}(W)$ is a neighborhood of c ;
- (b') W a neighborhood of $f(c) \Rightarrow \exists$ neighborhood V of c such that $f(V) \subset W$;
- (c) $(\forall \epsilon > 0) \exists \delta > 0 \ni d(x, c) < \delta \Rightarrow \rho(f(x), f(c)) < \epsilon$.

Proof. (a) \Rightarrow (b): If W is a neighborhood of $f(c)$, then $c \in f^{-1}(W)$, so assuming $x_n \rightarrow c$ we need only show that $x_n \in f^{-1}(W)$ ultimately (3.2.19). Indeed, $f(x_n) \rightarrow f(c) \in W$ by (a), therefore $f(x_n) \in W$ ultimately (3.2.19).

(b) \Rightarrow (b'): If W is a neighborhood of $f(c)$, then the set $V = f^{-1}(W)$ is a neighborhood of c by (b), and $f(V) = f(f^{-1}(W)) \subset W$.

(b') \Rightarrow (c): If $\epsilon > 0$ then the open ball $W = U_\epsilon(f(c))$ is a neighborhood of $f(c)$, so by (b') there exists a neighborhood V of c such that $f(V) \subset W$; any $\delta > 0$ such that $U_\delta(c) \subset V$ meets the requirements of (c).

(c) \Rightarrow (a): Assuming $x_n \rightarrow c$ we must show that $f(x_n) \rightarrow f(c)$. Given any $\epsilon > 0$, choose $\delta > 0$ as in (c); ultimately $d(x_n, c) < \delta$ and therefore $\rho(f(x_n), f(c)) < \epsilon$. \diamond

Condition (b) suggests the generalization to topological spaces:

3.4.3. *Definition.* Let X and Y be topological spaces, $f : X \rightarrow Y$, $c \in X$. We say that f is **continuous** at c if it satisfies condition (b) of 3.4.2. {It is equivalent to say that f satisfies condition (b').} If f is continuous at *every* point of X , we say simply that f is **continuous** on X (or is a *continuous mapping* of X into Y).

3.4.4. *Theorem.* Let X, Y, Z be topological spaces, let $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ be mappings, and let $x \in X$. If f is continuous at x , and g is continuous at $f(x)$, then the composite function $g \circ f : X \rightarrow Z$ is continuous at x .

Proof. Write $y = f(x)$ and $z = g(y) = g(f(x)) = (g \circ f)(x)$. If W is a neighborhood of $z = g(y)$ in Z , then $g^{-1}(W)$ is a neighborhood of y in Y , therefore $(g \circ f)^{-1}(W) = f^{-1}(g^{-1}(W))$ is a neighborhood of x in X . \diamond

3.4.5. *Theorem.* Let X and Y be topological spaces, $f : X \rightarrow Y$. The following conditions are equivalent:

- (a) f is continuous;
- (b) V open in $Y \Rightarrow f^{-1}(V)$ is open in X ;
- (c) B closed in $Y \Rightarrow f^{-1}(B)$ is closed in X .

Proof. The chain of implications (a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (b) \Rightarrow (a) is not the shortest possible, but each implication has a brief conceptual proof.

(a) \Rightarrow (b): Let V be an open set in Y . If $x \in f^{-1}(V)$ then $f(x) \in V$; since V is a neighborhood of $f(x)$ and f is continuous at x , $f^{-1}(V)$ is a neighborhood of x . Thus $f^{-1}(V)$ is a neighborhood of each of its points.

(b) \Rightarrow (c): If B is a closed set in Y , then $\complement B$ is an open set, therefore $f^{-1}(\complement B)$ is open by (b); thus $\complement f^{-1}(B) = f^{-1}(\complement B)$ is open, therefore $f^{-1}(B)$ is closed.

(c) \Rightarrow (b): The argument is similar to the preceding.

(b) \Rightarrow (a): If $x \in X$ and W is a neighborhood of $f(x)$, choose an open set V with $f(x) \in V \subset W$ (3.3.8); then $f^{-1}(V)$ is open by (b) and $x \in f^{-1}(V) \subset f^{-1}(W)$, therefore $f^{-1}(W)$ is a neighborhood of x . This shows that f is continuous at every $x \in X$. \diamond

3.4.6. Corollary. If $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ are continuous functions, then so is the composite function $g \circ f : X \rightarrow Z$.

Proof. If W is an open set in Z , then $g^{-1}(W)$ is open in Y , therefore $(g \circ f)^{-1}(W) = f^{-1}(g^{-1}(W))$ is open in X . \diamond

3.4.7. Definition. Topological spaces X and Y are said to be **homeomorphic** if there exists a bijection $f : X \rightarrow Y$ such that both f and f^{-1} are continuous; such a mapping is said to be a **homeomorphism** of X onto Y .

3.4.8. Example. The bijection $f : [-1, 1] \rightarrow \overline{\mathbb{R}}$ of Example 3.3.17 is a homeomorphism for the indicated metric topologies.

Exercises

1. For a function $f : X \rightarrow Y$ between topological spaces, the following conditions are equivalent: (a) f is continuous; (d) $f(\overline{A}) \subset \overline{f(A)}$ for every subset A of X .

{Hint: Use criterion (c) of 3.4.5.}

2. If X is a trivial topological space and Y is discrete (3.3.2), then every continuous function $f : X \rightarrow Y$ is a constant function.

{Hint: Each point of X has only one neighborhood.}

3. If Y is a discrete space then every continuous function $f : \mathbb{R} \rightarrow Y$ is constant.

{Hint: $f^{-1}(\{f(0)\})$ is both closed and open.}

4. Let (X, d) be a metric space, A a nonempty subset of X . For each $x \in X$, the nonnegative real number

$$\inf\{d(x, a) : a \in A\}$$

is called the *distance* from x to A and is denoted $d(x, A)$. The formula $f(x) = d(x, A)$ defines a function $f : X \rightarrow \mathbb{R}$.

(i) $f^{-1}(\{0\}) = \overline{A}$.

(ii) f is continuous; indeed, $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in X$.

{Hint: (ii) For all $a \in A$, $f(x) = d(x, A) \leq d(x, a) \leq d(x, y) + d(y, a)$.}

5. If A is a nonempty closed subset of a metric space (X, d) and if y is a point of X with $y \notin A$, then there exists a continuous function $g : X \rightarrow [0, 1]$ such that g is zero at precisely the points of A , and $g(y) = 1$.

{Hint: With f as in Exercise 4, define $g(x) = \min\{f(y)^{-1}f(x), 1\}$.}

6. Let (X, d) be a metric space and equip the product set $X \times X$ with the metric

$$D((x, y), (x', y')) = d(x, x') + d(y, y')$$

(cf. §3.1, Exercise 6, with $n = 2$ and $p = 1$).

- (i) $(x_n, y_n) \rightarrow (x, y)$ for $D \Leftrightarrow x_n \rightarrow x$ and $y_n \rightarrow y$ for d .
 (ii) d is a continuous function $X \times X \rightarrow \mathbb{R}$.
 {Hint: (ii) Cf. 3.2.4.}

7. Let A be a subset of a topological space X and let $\varphi_A : X \rightarrow \mathbb{R}$ be its characteristic function (1.3.4). Prove:

- (i) φ_A is continuous at a point $a \in A$ if and only if a is interior to A .
 (ii) The set of discontinuities of φ_A is the boundary ∂A of A (cf. §3.3, Exercise 6).

3.5. Limit of a Function

Intuitively, continuity of a function f at a point c means that if x approaches c , then $f(x)$ approaches $f(c)$; the concept of limit merely asks that if x approaches c , then $f(x)$ approaches *something*. A definition broad enough to encompass the one-sided limits and derivatives of functions of a real variable is so complex notationally as to require some preparation; the best place to start is to revisit the motivating example.¹

Consider a function $g : [a, b] \rightarrow \mathbb{R}$ and a point c , $a < c < b$. The function g is said to be *differentiable* at c if the difference-quotient

$$\frac{g(x) - g(c)}{x - c}$$

has a limit L in \mathbb{R} as x approaches c , in the sense that for sequences (x_n) in $[a, b]$,

$$x_n \rightarrow c, x_n \neq c \Rightarrow \frac{g(x_n) - g(c)}{x_n - c} \rightarrow L.$$

(Such an L is unique and is called the *derivative* of g at c .) Writing

$$f(x) = \frac{g(x) - g(c)}{x - c},$$

we have a function $f : B \rightarrow \mathbb{R}$ defined on the subset $B = [a, b] - \{c\}$ of $[a, b]$, and the condition for differentiability reads

$$x_n \rightarrow c, x_n \in B \Rightarrow f(x_n) \rightarrow L.$$

For *right-differentiability*, we require the existence of an $L \in \mathbb{R}$ such that

$$x_n \rightarrow c, x_n > c \Rightarrow f(x_n) \rightarrow L;$$

here, the approximating sequences are confined to the subset $A = (c, b]$ of B , and the condition now reads

$$x_n \rightarrow c, x_n \in A \Rightarrow f(x_n) \rightarrow L.$$

¹ First course, Chapter 8, §1.

For left-differentiability, where $x_n < c$ is required, we consider instead $A = [a, c)$.

Thus, initially we have a metric space $X = [a, b]$, but the function whose limit is under consideration is defined only on a subset B of X . The point c of X at which the limit is contemplated happens not to belong to the domain B of f ; it is, however, *adherent* to B (3.3.12) and to each of the subsets A considered above. The following diagram summarizes matters:

$$\begin{array}{ccc} & X & \\ & \cup & \\ & B & \xrightarrow{f} \mathbb{R} \\ & \cup & \\ c \in \overline{A} & \supset & A \end{array}$$

For right-differentiability we take $A = (c, b]$, for left-differentiability $A = [a, c)$, and for differentiability $A = B$.

With this example in mind, the following definition becomes digestible:

3.5.1. *Definition.* Let X and Y be metric spaces, B a subset of X , $f : B \rightarrow Y$ a function defined on B and taking values in Y . Let A be a subset of B and let $c \in \overline{A}$ (the closure of A in X). Schematically:

$$\begin{array}{ccc} & X & \\ & \cup & \\ & B & \xrightarrow{f} Y \\ & \cup & \\ c \in \overline{A} & \supset & A \end{array}$$

If there exists a point $y \in Y$ such that

$$x_n \rightarrow c, x_n \in A \Rightarrow f(x_n) \rightarrow y,$$

we say that f has **limit** y as x tends to c through values in A , written

$$\lim_{x \rightarrow c, x \in A} f(x) = y.$$

If $A = B$ we write simply

$$\lim_{x \rightarrow c} f(x) = y.$$

3.5.2. *Remarks.* (i) The existence or nonexistence of y is entirely determined by the restriction $f|_A$ of f to A .

(ii) If such a y exists, then it is unique (3.2.1).

(iii) If $c \in B$ one can drop down to B and dispense with X . In this case $f(c)$ is defined, but if $c \notin A$ then $f(c)$ is irrelevant to the existence of y .

(iv) If $c \in A$ then $f(c)$ is defined and the notion of limit brings nothing new to the table; it is just continuity of $f|_A$ at c :

3.5.3. Theorem. *With notations as in 3.5.1, suppose also that $c \in A$. Then:*

$$\exists \lim_{x \rightarrow c, x \in A} f(x) \Leftrightarrow f|A \text{ is continuous at } c,$$

and in this case

$$\lim_{x \rightarrow c, x \in A} f(x) = f(c).$$

Proof. \Rightarrow : Let $y \in Y$ be the limit whose existence is assumed. Since $c \in A$ we are free to take the constant sequence $x_n = c$ in 3.5.1, therefore $f(c) = y$. Then $f(x_n) \rightarrow y = f(c)$ for every sequence $x_n \in A$ converging to c , in other words $f|A$ is continuous at c (where A is regarded as a metric space with the metric $d|A \times A$ it inherits from X).

\Leftarrow : By assumption, $f(x_n) \rightarrow f(c)$ for every sequence $x_n \in A$ with $x_n \rightarrow c$, so $f(c)$ meets the requirements for y in 3.5.1. \diamond

There are also ϵ, δ and neighborhood formulations of limit:

3.5.4. Theorem. *With notations as in 3.5.1, let d be the metric on X , and ρ the metric on Y . Let $y \in Y$. The following conditions are equivalent:*

- (a) $\lim_{x \rightarrow c, x \in A} f(x)$ exists and is equal to y ;
- (b) for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$x \in A, d(x, c) < \delta \Rightarrow \rho(f(x), y) < \epsilon;$$

- (c) for every neighborhood W of y in Y , there exists a neighborhood V of c in X such that $f(V \cap A) \subset W$.

Proof. (a) \Rightarrow (b): If, on the contrary, there is an $\epsilon > 0$ for which no suitable δ exists, then for each positive integer n there exists a point $x_n \in A$ such that $d(x_n, c) < 1/n$ but $\rho(f(x_n), y) \geq \epsilon$. Then $x_n \rightarrow c$ but $f(x_n) \not\rightarrow y$, contrary to (a).

(b) \Rightarrow (c): Let W be a neighborhood of y and choose $\epsilon > 0$ so that W contains the open ball $U_\epsilon(y)$ (3.2.12). For this ϵ , choose δ as in (b); then the open ball $U_\delta(c)$ meets the requirements for V .

(c) \Rightarrow (a): Let $x_n \in A$, $x_n \rightarrow c$. Given any $\epsilon > 0$, let $W = U_\epsilon(y)$ and choose V as in (c); ultimately $x_n \in V$, therefore $f(x_n) \in W$, that is, $\rho(f(x_n), y) < \epsilon$. This shows that $f(x_n) \rightarrow y$. \diamond

For functions of a real variable, the approach to c can be from either one or both sides. In the following definition, we restrict attention to the case that c is approachable through an adjacent interval from one or both sides; this covers the case of derivatives and is adequate for all of our applications:

3.5.5. Definition. Suppose $X = \mathbb{R}$ in 3.5.1, and suppose

$$\exists \lim_{x \rightarrow c, x \in A} f(x) = y.$$

(i) If B is a 'deleted neighborhood' of c , that is,

$$B \supset (c - r, c) \cup (c, c + r) \quad \text{for some } r > 0,$$

and if

$$A = B \cap (\mathbb{R} - \{c\}) = \{x \in B : x \neq c\},$$

then we write instead

$$\lim_{x \rightarrow c, x \neq c} f(x) = y.$$

The possibility that $c \in B$ (so that B is a neighborhood of c) is not ruled out; it just does not figure in the definition either here, or in the following definitions (ii) and (iii).

(ii) If B is a 'deleted right neighborhood' of c , that is,

$$B \supset (c, c + r) \quad \text{for some } r > 0,$$

and if

$$A = B \cap (c, +\infty) = \{x \in B : x > c\},$$

we write instead

$$\lim_{x \rightarrow c, x > c} f(x) = y \quad \text{or} \quad \lim_{x \rightarrow c+} f(x) = y,$$

and we say that f has **right limit** y at c , expressed concisely as $f(c+) = y$.

(iii) If B is a 'deleted left neighborhood' of c , that is,

$$B \supset (c - r, c) \quad \text{for some } r > 0,$$

and if

$$A = B \cap (-\infty, c) = \{x \in B : x < c\},$$

we write instead

$$\lim_{x \rightarrow c, x < c} f(x) = y \quad \text{or} \quad \lim_{x \rightarrow c-} f(x) = y,$$

and we say that f has **left limit** y at c , briefly $f(c-) = y$.

The 'two-sided' limit defined in (i) exists if and only if both of the one-sided limits in (ii), (iii) exist and are equal to each other:

3.5.6. Theorem. *Suppose $B \subset \mathbb{R}$ and $f : B \rightarrow Y$, where Y is a metric space, and let $c \in \mathbb{R}$. The following conditions are equivalent:*

- (a) $\lim_{x \rightarrow c, x \neq c} f(x)$ exists;
- (b) $f(c-), f(c+)$ exist and are equal.

When this is the case, all three numbers are equal to each other:

$$\lim_{x \rightarrow c, x \neq c} f(x) = f(c-) = f(c+).$$

Proof. Note first that the requirements on the domain of f are consistent in (a) and (b). The proof that (a) implies (b) with equality of all three numbers is obvious. The crux of the proof of (b) \Rightarrow (a) is that if $x_n \in B$, $x_n \neq c$, $x_n \rightarrow c$, then either $x_n < c$ ultimately, or $x_n > c$ ultimately, or (x_n) decomposes into two subsequences, one each of the preceding two types. \diamond

The application of the foregoing to derivatives formalizes the discussion in the introduction. The derivative (two-sided or one-sided) of a function f is calculated only at points c belonging to an interval contained in the domain of f , so we need only consider the core case of a function defined on a nondegenerate closed interval:

3.5.7. *Definition.* Suppose $f : [a, b] \rightarrow \mathbb{R}$, where $a < b$, and let $c \in [a, b]$. Let $B = [a, b] - \{c\}$ and consider the difference-quotient function $B \rightarrow \mathbb{R}$ defined by

$$x \mapsto \frac{f(x) - f(c)}{x - c} \quad (x \in B).$$

We are going to apply the machinery of 3.5.5 to this difference-quotient function:

(i) If $a < c < b$ and if

$$\exists \lim_{x \rightarrow c, x \neq c} \frac{f(x) - f(c)}{x - c} \quad \text{in } \mathbb{R},$$

then f is said to be **differentiable** at c ; the limit is then denoted $f'(c)$, called the **derivative** of f at c :

$$f'(c) = \lim_{x \rightarrow c, x \neq c} \frac{f(x) - f(c)}{x - c}.$$

(ii) If $a \leq c < b$ and if

$$\exists \lim_{x \rightarrow c, x > c} \frac{f(x) - f(c)}{x - c} \quad \text{in } \mathbb{R},$$

then f is said to be **right-differentiable** at c ; this limit is called the **right-derivative** of f at c and is denoted $f'_r(c)$:

$$f'_r(c) = \lim_{x \rightarrow c, x > c} \frac{f(x) - f(c)}{x - c}.$$

(iii) If $a < c \leq b$ then **left-differentiability** and the **left derivative** $f'_l(c) \in \mathbb{R}$ are defined analogously,

$$f'_l(c) = \lim_{x \rightarrow c, x < c} \frac{f(x) - f(c)}{x - c}.$$

3.5.8. *Theorem.* With notations as in 3.5.7, suppose $a < c < b$. The following conditions are equivalent:

(a) f is differentiable at c ;

(b) f is both left- and right-differentiable at c , and $f'_l(c) = f'_r(c)$.

When this is the case, $f'(c) = f'_l(c) = f'_r(c)$.

Proof. This is immediate from 3.5.6 (applied to the difference-quotient function) and the definitions. \diamond

Exercises

1. In the definition of limit (3.5.1), X and Y need only be metrizable topological spaces, since the notions of sequential convergence and closure do not change when one passes to an equivalent metric (cf. 3.2.19).

(i) Let X and Y be topological spaces and, as in the diagram of 3.5.1, suppose that $A \subset B \subset X$, $f: B \rightarrow Y$ and $c \in \overline{A}$. Let $y \in Y$. Guided by condition (c) of 3.5.4, we say that f has *limit* y as x tends to c through values in A , written

$$\lim_{x \rightarrow c, x \in A} f(x) = y \quad (\text{briefly } \lim_{x \rightarrow c} f(x) = y, \text{ when } A = B),$$

if, for every neighborhood W of y in Y , there exists a neighborhood V of c in X such that $f(V \cap A) \subset W$. {CAUTION: If Y has the trivial topology (3.3.2, (ii)) then $\lim_{x \rightarrow c, x \in A} f(x) = y$ for every $c \in \overline{A}$ and every $y \in Y$.}

(ii) In the context of (i), the limit y will be uniquely determined by f if the following property holds: if y and y' are distinct elements of Y , then there exist neighborhoods W of y and W' of y' such that $W \cap W' = \emptyset$. {Such a topological space is called a *Hausdorff space* (or a *separated space*, since distinct points can be separated by means of disjoint open sets).}

(iii) In the context of (i), suppose $c \in A$ and equip A with the relative topology induced by the topology of X (§3.3, Exercise 7). Then the restriction $f|_A$ is continuous at c if and only if $\lim_{x \rightarrow c, x \in A} f(x) = f(c)$.

{Hint: The sets $V \cap A$ described in (i) are neighborhoods of c in A for the relative topology.}

2. With notations as in the definition of limit in 3.5.1 (X and Y metric spaces), if $c \in A$ then f has a limit as $x \in A$ tends to c if and only if $f|_A$ is continuous at c (3.5.3).

On the other hand, if $c \notin A$ then the following conditions are equivalent:

(a) $\lim_{x \rightarrow c, x \in A} f(x)$ exists;

(b) there exists a function $g: A \cup \{c\} \rightarrow Y$ such that $g(x) = f(x)$ for all $x \in A$ and g is continuous at c (for the relative topology on $A \cup \{c\}$ induced by the topology of X).

When this is the case, necessarily $\lim_{x \rightarrow c, x \in A} f(x) = g(c)$.

The equivalence of (a) and (b) (under the assumption $c \notin A$) remains valid in the topological setting of Exercise 1, (i), and $g(c)$ is then a determination of $\lim_{x \rightarrow c, x \in A} f(x)$ (but see the "CAUTION" in Exercise 1).

{Hint: (a) \Rightarrow (b): If y is the assumed limit, define $g : A \cup \{c\} \rightarrow Y$ by the formulas $g(c) = y$ and $g(x) = f(x)$ for all $x \in A$. If W is a neighborhood of $y = g(c)$, and V is a neighborhood of c in X with $f(V \cap A) \subset W$, contemplate $V \cap (A \cup \{c\}) = (V \cap A) \cup \{c\}$.

(b) \Rightarrow (a): If W is a neighborhood of $g(c)$ in Y , and V is a neighborhood of x in X such that $g(V \cap (A \cup \{c\})) \subset W$, then $f(V \cap A) \subset W$.}

CHAPTER 4

Lebesgue Integral

- §4.1. Measurable functions
- §4.2. a.e.
- §4.3. Integrable simple functions
- §4.4. Integrable functions
- §4.5. Monotone convergence theorem, Fatou's lemma
- §4.6. Monotone classes
- §4.7. Indefinite integrals
- §4.8. Finite signed measures

Chapter 2 lays the foundation for the Lebesgue integral. What comes first is the concept of outer measure, a natural extension of interval length; the class of measurable sets is then defined in terms of the outer measure, guided by the need for additivity of measure as a set function. Outer measure provides the raw numbers; measurability provides the 'smooth' sets on which the numbers are well-behaved. Such well-behavior is codified in the concept of measure space (2.4.12).

The context for general integration theory is a measure space (X, \mathcal{S}, μ) , the Lebesgue integral being obtained by specializing to the measure space $(\mathbb{R}, \mathcal{M}, \lambda)$ of 2.4.13. The measure μ assigns numbers (in $[0, +\infty]$) to the sets of \mathcal{S} ; the integral assigns numbers (in \mathbb{R}) to certain functions $f : X \rightarrow \mathbb{R}$ describable in terms of the sets of \mathcal{S} . The natural place to start is with characteristic functions: if $f = \varphi_E$, where $E \in \mathcal{S}$ has finite measure, the integral of f is defined to be $\mu(E)$. The next step is to consider finite linear combinations $f = c_1\varphi_{E_1} + \dots + c_n\varphi_{E_n}$ of such characteristic functions, and, finally, suitable limits $f = \lim f_k$ of sequences of such linear combinations. The functions (called integrable) that make it to the finish line will, in particular, have been built out of the sets of \mathcal{S} .

Thinking of \mathcal{S} as the set of 'measurable' sets of the theory, functions that can be 'built out of' the sets of \mathcal{S} are said to be 'measurable' with respect to the σ -algebra \mathcal{S} ; the chapter begins with a study of such functions, without regard to measures that may be defined on \mathcal{S} .

4.1. Measurable Functions

4.1.1. *Definition.* A **measurable space** (or **Borel space**) is a pair (X, \mathcal{S}) , where X is a set and \mathcal{S} is a σ -algebra of subsets of X (2.4.1). The sets in \mathcal{S} are called the **measurable** subsets of X . (Despite the word 'measurable', the concept is purely set-theoretic; there is no measure in view.)

4.1.2. *Examples.* (i) For any set X , the pairs $(X, \{\emptyset, X\})$ and $(X, \mathcal{P}(X))$ are measurable spaces.

(ii) If \mathcal{B} is the class of Borel sets in \mathbb{R} , then the pair $(\mathbb{R}, \mathcal{B})$ is a measurable space (2.4.5). More generally, if X is a topological space and $\mathcal{B}(X)$ is the σ -algebra generated (2.4.4) by the class of open sets in X , then $(X, \mathcal{B}(X))$ is a measurable space; the sets in $\mathcal{B}(X)$ are called the *Borel sets* of the topological space X .

(iii) If \mathcal{M} is the class of Lebesgue-measurable sets in \mathbb{R} , then $(\mathbb{R}, \mathcal{M})$ is a measurable space (2.4.3). More generally, if ρ is an outer measure on a set X and if $\mathcal{M}(\rho)$ is the class of all ρ -measurable sets in X , then $(X, \mathcal{M}(\rho))$ is a measurable space (2.4.3).

For the rest of the section, (X, \mathcal{S}) denotes a fixed measurable space.

4.1.3. *Definition.* A function $f : X \rightarrow \mathbb{R}$ is said to be **measurable** (with respect to \mathcal{S}) if the inverse image of every Borel set of \mathbb{R} under f is measurable, that is,

$$B \text{ a Borel set in } \mathbb{R} \Rightarrow f^{-1}(B) \in \mathcal{S}.$$

4.1.4. *Remarks.* (i) We have in the preceding definition a mapping $f : (X, \mathcal{S}) \rightarrow (\mathbb{R}, \mathcal{B})$ between Borel spaces; measurability means that $f^{-1}(\mathcal{B}) \subset \mathcal{S}$, expressed by saying that f is a *morphism* for the given Borel structures of X and \mathbb{R} . Note the analogy with mappings $f : Y \rightarrow Z$ between topological spaces, the condition for continuity being $f^{-1}(\mathcal{O}_Z) \subset \mathcal{O}_Y$, where \mathcal{O} denotes the class of open sets; one says that the continuous mappings $f : Y \rightarrow Z$ are the morphisms for the topological structures of Y and Z .

(ii) One could allow measurable functions to have infinite values (Exercise 1); however, sums of such functions are in general only partially defined, since infinite values of opposite sign cannot be added (1.15.4). The restriction to finite-valued functions simplifies the algebra of measurable functions and is adequate for all of our applications.

4.1.5. *Examples.* (i) Every constant real function on X is measurable; for, if $f(x) \equiv c$ and B is a Borel set in \mathbb{R} , then $f^{-1}(B)$ is either X or \emptyset according as $c \in B$ or $c \notin B$.

(ii) For a subset E of X , the characteristic function $f = \varphi_E$ of E is measurable $\Leftrightarrow E \in \mathcal{S}$; for, if B is a Borel set in \mathbb{R} , then $f^{-1}(B)$ is one of the sets $X, \emptyset, E, \mathbb{C}E$, according as $0, 1$ do or do not belong to B .

(ii) A real-valued function f on \mathbb{R} is said to be *Lebesgue-measurable* if it is measurable with respect to the measurable space $(\mathbb{R}, \mathcal{M})$ of Example 4.1.2, (iii).

4.1.6. Theorem. Let \mathcal{E} be a set of Borel sets in \mathbb{R} that generates the σ -algebra \mathcal{B} of all Borel sets (2.4.4). The following conditions on a function $f : X \rightarrow \mathbb{R}$ are equivalent:

- (a) f is measurable;
- (b) $f^{-1}(A) \in \mathcal{S}$ for all $A \in \mathcal{E}$.

Proof. (b) \Rightarrow (a): Let $\mathcal{C} = \{C \subset \mathbb{R} : f^{-1}(C) \in \mathcal{S}\}$. By assumption, $\mathcal{E} \subset \mathcal{C}$; the problem is to show that $\mathcal{B} \subset \mathcal{C}$, and for this it suffices to show that \mathcal{C} is a σ -algebra (2.4.4). Let us verify the three conditions of 2.4.1.

From $f^{-1}(\emptyset) = \emptyset \in \mathcal{S}$ we see that $\emptyset \in \mathcal{C}$. If $f^{-1}(C) \in \mathcal{S}$ then also $f^{-1}(\mathbb{C}C) = \mathbb{C}f^{-1}(C) \in \mathcal{S}$, thus \mathcal{C} is closed under complementation. If (C_n) is a sequence of sets in \mathbb{R} such that $f^{-1}(C_n) \in \mathcal{C}$ for all n , then $f^{-1}(\bigcup C_n) = \bigcup f^{-1}(C_n) \in \mathcal{S}$, thus \mathcal{C} is closed under countable unions. \diamond

4.1.7. Examples. (i) With notations as in 4.1.6, we may take \mathcal{E} to be the set of all open intervals in \mathbb{R} (2.4.5), or the set of all closed intervals (cf. §2.4, Exercise 1). Indeed, \mathcal{E} can be the set of all intervals of each of the eight types (a, b) , $[a, b)$, $(a, b]$, $[a, b]$, $(c, +\infty)$, $[c, +\infty)$, $(-\infty, c)$, $(-\infty, c]$; the point is that the σ -algebra generated by each type of interval contains all the other types. For example, if $a < b$ and if (b_n) is a sequence with $a < b_n < b$ and $b_n \rightarrow a$, then the formulas

$$(a, b) = \bigcup_{n=1}^{\infty} [b_n, b), \quad [a, b) = \bigcap_{n=1}^{\infty} (a - 1/n, b)$$

show that the intervals $[a, b)$ generate the same σ -algebra as the intervals (a, b) (namely, the σ -algebra \mathcal{B} of all Borel sets).

(ii) In (i) we need only consider intervals with rational endpoints. In particular, \mathcal{B} is generated by a countable set of intervals.

4.1.8. Corollary. The following conditions on a function $f : X \rightarrow \mathbb{R}$ are equivalent:

- (a) f is measurable;
- (b) $\{x : f(x) > c\} \in \mathcal{S}$ for every real (or every rational) number c ;
- (c) $\{x : f(x) \leq c\} \in \mathcal{S}$ for every real (or every rational) number c .

Proof. The conditions (b) and (c) are equivalent to each other by complementation, and (b) is equivalent to (a) by 4.1.6 and the formula $\{x : f(x) > c\} = f^{-1}((c, +\infty))$. \diamond

Similar statements hold with $<$ and \geq in place of $>$ and \leq .

The measurable functions on X form a linear subspace of the vector space of all real-valued functions on X :

4.1.9. Theorem. If f and g are measurable functions on X and if $a \in \mathbb{R}$, then the functions $f + g$ and af are also measurable.

Proof. Concerning af : If $a = 0$ then af is the constant function 0. If $a > 0$ then, for every $c \in \mathbb{R}$,

$$\{x : (af)(x) < c\} = \{x : f(x) < c/a\} \in \mathcal{S},$$

whereas if $a < 0$ then

$$\{x : (af)(x) < c\} = \{x : f(x) > c/a\} \in \mathcal{S}$$

for all $c \in \mathbb{R}$.

Concerning $f + g$: If $c \in \mathbb{R}$ then

$$\begin{aligned} f(x) + g(x) < c &\Leftrightarrow f(x) < c - g(x) \\ &\Leftrightarrow \exists r \in \mathbb{Q} \ni f(x) < r < c - g(x) \\ &\Leftrightarrow \exists r \in \mathbb{Q} \ni f(x) < r \text{ \& } g(x) < c - r, \end{aligned}$$

thus

$$\{x : (f + g)(x) < c\} = \bigcup_{r \in \mathbb{Q}} \{x : f(x) < r\} \cap \{x : g(x) < c - r\},$$

which is a countable union of measurable sets. \diamond

4.1.10. Definition. A function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is said to be a **Borel function** if it is measurable with respect to the σ -algebra \mathcal{B} of Borel sets, that is,

$$B \in \mathcal{B} \Rightarrow \varphi^{-1}(B) \in \mathcal{B}.$$

4.1.11. Example. Every continuous function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is Borel; for, if U is any open set in \mathbb{R} then $\varphi^{-1}(U)$ is also open, hence Borel (2.4.11), therefore φ is Borel by 4.1.6.

4.1.12. Lemma. If $f : X \rightarrow \mathbb{R}$ is measurable and $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is Borel,

$$X \xrightarrow{f} \mathbb{R} \xrightarrow{\varphi} \mathbb{R},$$

then the composite function $\varphi \circ f$ is measurable.

Proof. For every Borel set B in \mathbb{R} , $\varphi^{-1}(B)$ is Borel by assumption, therefore $(\varphi \circ f)^{-1}(B) = f^{-1}(\varphi^{-1}(B)) \in \mathcal{S}$ by the measurability of f . \diamond

4.1.13. Theorem. If f and g are measurable functions on X and if $\alpha > 0$, then the following functions on X are also measurable: $|f|^\alpha$, $f \cup g$, $f \cap g$ and fg .

Proof. Here $(f \cup g)(x) = \max\{f(x), g(x)\}$, $(f \cap g)(x) = \min\{f(x), g(x)\}$. The function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\varphi(t) = |t|^\alpha$ is continuous, therefore Borel (4.1.11); the composite function $x \mapsto \varphi(f(x)) = |f(x)|^\alpha =$

$|f|^\alpha(x)$ is therefore measurable by the lemma. In particular $|f|$ and f^2 are measurable, therefore the functions

$$\begin{aligned} f \cup g &= \frac{1}{2}\{f + g + |f - g|\} \\ f \cap g &= \frac{1}{2}\{f + g - |f - g|\} \\ fg &= \frac{1}{4}\{(f + g)^2 - (f - g)^2\} \end{aligned}$$

are measurable by 4.1.9. \diamond

By the formulas in the preceding proof,

$$(f \cup g) + (f \cap g) = f + g, \quad (f \cup g) - (f \cap g) = |f - g|;$$

putting $g = 0$, we have

$$f = (f \cup 0) + (f \cap 0), \quad |f| = (f \cup 0) - (f \cap 0).$$

4.1.14. *Definition.* For any function $f : X \rightarrow \mathbb{R}$, we write $f^+ = f \cup 0$, $f^- = -(f \cap 0)$, that is,

$$f^+(x) = \max\{f(x), 0\}, \quad f^-(x) = -\min\{f(x), 0\};$$

thus $f^+ \geq 0$, $f^- \geq 0$, $f^+ f^- = 0$ and, by the preceding remarks,

$$f = f^+ - f^-, \quad |f| = f^+ + f^-.$$

One calls f^+ the *positive part* of f , and $-f^- = f \cap 0$ the *negative part* of f (thus f is the sum of its positive and negative parts).

4.1.15. *Theorem.* For a function $f : X \rightarrow \mathbb{R}$,

$$f \text{ is measurable} \Leftrightarrow f^+ \text{ and } f^- \text{ are measurable.}$$

Proof. \Rightarrow : Immediate from 4.1.13.

\Leftarrow : If f^+, f^- are measurable, then so is their difference f (4.1.9). \diamond

We now look at preservation of measurability under passage to limits. In the next definitions, X can be any set (the σ -algebra \mathcal{S} is irrelevant) and the functions on X can have infinite values:

4.1.16. *Definition.* Let $f_n : X \rightarrow \overline{\mathbb{R}}$ ($n = 1, 2, 3, \dots$) be a sequence of extended real-valued functions on X . The functions

$$\sup_n f_n : X \rightarrow \overline{\mathbb{R}}, \quad \inf_n f_n : X \rightarrow \overline{\mathbb{R}}$$

are the pointwise supremum and infimum of the sequence (f_n) , defined by the formulas

$$(\sup_n f_n)(x) = \sup_n f_n(x), \quad (\inf_n f_n)(x) = \inf_n f_n(x).$$

The pointwise \limsup and \liminf of the sequences are then defined by the

formulas

$$\begin{aligned}\limsup f_n &= \inf_{n \geq 1} \left(\sup_{k \geq n} f_k \right), \\ \liminf f_n &= \sup_{n \geq 1} \left(\inf_{k \geq n} f_k \right).\end{aligned}$$

4.1.17. *Remarks.* (i) For each $x \in X$,

$$\begin{aligned}(\limsup f_n)(x) &= \limsup f_n(x), \\ (\liminf f_n)(x) &= \liminf f_n(x),\end{aligned}$$

where the right sides are defined as in 1.16.1.

(ii) $\liminf f_n \leq \limsup f_n$ at every point of X (1.16.2).

(iii) $(\liminf f_n)(x) = (\limsup f_n)(x) \Leftrightarrow (f_n(x))$ is convergent in $\overline{\mathbb{R}}$ (1.16.8).

4.1.18. *Definition.* With notations as in 4.1.16, the sequence (f_n) is said to be **pointwise convergent** on X if $(f_n(x))$ is convergent in $\overline{\mathbb{R}}$ for every $x \in X$, that is, if $\liminf f_n = \limsup f_n$; when this is the case, the function $f : X \rightarrow \overline{\mathbb{R}}$ defined by

$$f(x) = \lim f_n(x) \quad (\forall x \in X)$$

is called the **pointwise limit** of the sequence (f_n) , we write $f = \lim f_n = \liminf f_n = \limsup f_n$ and we say that $f_n \rightarrow f$ pointwise on X .

4.1.19. *Theorem.* If $f_n : X \rightarrow \mathbb{R}$ is a sequence of measurable functions that is pointwise bounded above on X , in the sense that

$$\sup_n f_n(x) < +\infty \quad \text{for all } x \in X,$$

then the function $\sup f_n$ is also measurable.

Proof. Write $g = \sup_n f_n$; by hypothesis, g is real-valued. If $c \in \mathbb{R}$, then

$$g(x) \leq c \Leftrightarrow f_n(x) \leq c \text{ for all } n;$$

thus, the set

$$\{x : g(x) \leq c\} = \bigcap_{n=1}^{\infty} \{x : f_n(x) \leq c\}$$

is the intersection of a sequence of measurable sets, therefore g is measurable by 4.1.8. \diamond

4.1.20. *Corollary.* Let $f_n : X \rightarrow \mathbb{R}$ be a sequence of measurable functions.

(i) If the sequence (f_n) is pointwise bounded below on X , then $\inf_n f_n$ is measurable.

(ii) If (f_n) is pointwise bounded, then $\liminf f_n$ and $\limsup f_n$ are measurable.

(iii) If f is a real-valued function on X such that $f_n \rightarrow f$ pointwise, then f is measurable.

Proof. (i) $\inf f_n = -\sup(-f_n)$ is measurable by the theorem.

(ii) This is clear from the theorem, (i) and the definitions.

(iii) Since (f_n) is pointwise convergent it is pointwise bounded, therefore $f = \lim f_n = \liminf f_n$ is measurable by (ii). \diamond

4.1.21. *Definition.* A function $f : X \rightarrow \mathbb{R}$ is said to be **simple** (relative to \mathcal{S}) if (i) f is measurable, and (ii) $f(X)$ is finite.

The simple functions form a linear subspace of the vector space of all measurable functions, and more:

4.1.22. *Theorem.* If f and g are simple functions on X and if $a \in \mathbb{R}$, then the following functions are also simple: $f + g$, af , fg , $|f|$, $f \cup g$, $f \cap g$, f^+ and f^- .

Proof. The functions in question are all measurable (4.1.9, 4.1.13, 4.1.15) and it is clear that each has only finitely many values. \diamond

Here are two useful characterizations of simplicity:

4.1.23. *Theorem.* For a function $f : X \rightarrow \mathbb{R}$, the following conditions are equivalent:

- (a) f is simple;
- (b) f has only finitely many values, each assumed on a measurable set;
- (c) f is a linear combination of characteristic functions of measurable sets.

Proof. (a) \Rightarrow (b): If c is one of the values of f , then $f^{-1}(\{c\}) \in \mathcal{S}$ because $\{c\}$ is a Borel set and f is measurable.

(b) \Rightarrow (c): If c_1, \dots, c_n are the distinct values of f , let $E_i = \{x : f(x) = c_i\}$ ($i = 1, \dots, n$); the E_i are pairwise disjoint sets, measurable by hypothesis, with union X , and it is clear that $f = \sum_{i=1}^n c_i \varphi_{E_i}$.

(c) \Rightarrow (a): If $f = \sum_{j=1}^m a_j \varphi_{F_j}$ with $a_j \in \mathbb{R}$ and $F_j \in \mathcal{S}$ for $j = 1, \dots, m$, then f is measurable by 4.1.5, (ii) and 4.1.9. If $x \in X$ then $f(x)$ is the sum of those a_j for which $x \in F_j$, thus f has at most 2^m values. \diamond

A pointwise limit of simple functions is obviously measurable (4.1.20); we are going to show, conversely, that every measurable function is the pointwise limit of a sequence of simple functions.

4.1.24. *Lemma.* If $f : X \rightarrow \mathbb{R}$ is bounded and measurable, and if $f \geq 0$ on X , then there exists a simple function g such that $0 \leq g \leq f$ and $\|f - g\|_\infty \leq \frac{1}{2} \|f\|_\infty$.

Proof. Let $a = \frac{1}{2}\|f\|_\infty$, $E = \{x : f(x) > a\}$ and $g = a\varphi_E$. \diamond

4.1.25. Theorem. If $f : X \rightarrow \mathbb{R}$ is bounded and measurable, then there exists a sequence (f_n) of simple functions such that $\|f_n - f\|_\infty \rightarrow 0$. If, moreover, $f \geq 0$, we can suppose that $0 \leq f_1 \leq f_2 \leq f_3 \leq \dots$ (concisely, $0 \leq f_n \uparrow f$ pointwise on X).

Proof. Writing $f = f^+ - f^-$, we need only consider the case that $f \geq 0$ (4.1.15). Let $c = \|f\|_\infty$. By the lemma, there exists a simple function g_1 such that $0 \leq g_1 \leq f$ and $\|f - g_1\|_\infty \leq \frac{1}{2}c$. Applying the lemma to $f - g_1$, there exists a simple function g_2 such that $0 \leq g_2 \leq f - g_1$ and $\|(f - g_1) - g_2\|_\infty \leq \frac{1}{2}\|f - g_1\|_\infty \leq (\frac{1}{2})^2 c$. Continuing, we construct recursively a sequence (g_n) of nonnegative simple functions such that

$$0 \leq g_1 + \dots + g_n \leq f \quad \text{and} \quad \|f - (g_1 + \dots + g_n)\|_\infty \leq (\frac{1}{2})^n c,$$

thus the functions $f_n = g_1 + \dots + g_n$ meet the requirements of the theorem. \diamond

The assumption of boundedness can be dropped, at the cost of weakening uniform convergence to pointwise convergence:

4.1.26. Theorem. For a function $f : X \rightarrow \mathbb{R}$, the following conditions are equivalent:

- (a) f is measurable;
- (b) there exists a sequence (f_n) of simple functions such that $f_n \rightarrow f$ pointwise on X .

If, moreover, $f \geq 0$, we can suppose that $0 \leq f_n \uparrow f$.

Proof. (b) \Rightarrow (a): This is immediate from 4.1.20, (iii).

(a) \Rightarrow (b): Writing $f = f^+ - f^-$, we can suppose that $f \geq 0$. For every positive integer n , let $g_n = f \cap (n1)$; g_n is measurable (4.1.13) and bounded (by n), and $0 \leq g_n \uparrow f$ pointwise on X . By the preceding theorem there exists, for each n , a simple function h_n such that $0 \leq h_n \leq g_n$ and $\|g_n - h_n\|_\infty \leq 1/n$; since $g_n \rightarrow f$ pointwise, it follows that $h_n \rightarrow f$ pointwise. The functions $f_n = h_1 \cup \dots \cup h_n$ are simple (4.1.22), $f_1 \leq f_2 \leq f_3 \leq \dots$, and

$$0 \leq h_n \leq f_n \leq g_1 \cup \dots \cup g_n = g_n \leq f;$$

since $h_n \rightarrow f$ we infer that $f_n \rightarrow f$ pointwise. \diamond

Exercises

1. If (X, \mathcal{S}) is a measurable space, define measurability for an extended real-valued function $f : X \rightarrow \overline{\mathbb{R}}$.

{Hint: First describe the Borel sets of $\overline{\mathbb{R}}$ regarded as a topological space (cf. 3.3.17).}

2. If $|f|$ is measurable, does it follow that f is measurable?

3. If f and g are measurable real-valued functions on a measurable space (X, \mathcal{S}) , then the function h defined by

$$h(x) = \begin{cases} \frac{f(x)}{g(x)} & \text{when } g(x) \neq 0 \\ 0 & \text{when } g(x) = 0 \end{cases}$$

is also measurable. More generally, if r is any measurable real-valued function on X , one can require that $h(x) = r(x)$ whenever $g(x) = 0$.

{Hint: Consider first the case that $g \geq 0$ and $f = 1$. For the general case, contemplate the formula $f(x)/g(x) = (fg)(x)[1/g^2(x)]$.}

4. A measurable real-valued function f can be written as $f = u|f|$, where u is a measurable function such that $u(x) = \pm 1$ for every x .

{Hint: In Exercise 3, consider $g = |f|$ and $r = 1$.}

4.2. a.e.

To the measurable space (X, \mathcal{S}) of the preceding section, we now add a measure μ : For the rest of the section, (X, \mathcal{S}, μ) denotes a fixed measure space (2.4.12).

The abbreviation 'a.e.' of the title stands for 'almost everywhere'; the definition is based on the idea of 'negligible set', which, by an amusing twist of terminology, is the single most important concept in the theory of the Lebesgue integral:

4.2.1. *Definition.* A **null set** (or μ -null set) is a measurable set of measure zero, that is, a set $E \in \mathcal{S}$ such that $\mu(E) = 0$. A set $A \subset X$ is said to be **negligible** (or μ -negligible) if $A \subset E$ for some null set E .

4.2.2. *Remarks.* (i) Every null set is negligible; a negligible set is null if and only if it is measurable.

(ii) The empty set \emptyset is null. If (E_n) is a sequence of null sets, then its union $\bigcup_{n=1}^{\infty} E_n$ is null. If E is null and F is measurable, then $E \cap F$ is null.

(iii) The union of a sequence of negligible sets is negligible. Every subset of a negligible set is negligible.

(iv) For the Lebesgue measure space $(\mathbb{R}, \mathcal{M}, \lambda)$, null \Leftrightarrow negligible. Such measure spaces are called *complete*. More generally:

(v) If ρ is an outer measure on X and $\mathcal{M} = \mathcal{M}(\rho)$ is the set of all ρ -measurable sets, then $(X, \mathcal{M}, \rho|_{\mathcal{M}})$ is a complete measure space (cf. 2.4.13).

(vi) If $\mathcal{B} = \mathcal{B}(\mathbb{R})$ is the σ -algebra of Borel sets in \mathbb{R} and λ is Lebesgue measure, then the measure space $(\mathbb{R}, \mathcal{B}, \lambda|_{\mathcal{B}})$ is not complete (cf. §2.4, Exer. 2).

4.2.3. *Definition.* Let S be a subset of X and, for each $x \in S$ let $P(x)$ be a proposition (typically a statement about x). We say $P(x)$ is true a.e. in S (usually abbreviated to ' $P(x)$ a.e. in S '), relative to the measure μ , if the set $A = \{x \in S : P(x) \text{ false}\}$ is negligible, that is, $A \subset E \in \mathcal{S}$ with $\mu(E) = 0$; in other words, there exists a null set E such that

$$x \in S - E \Rightarrow P(x) \text{ true.}$$

This is also expressed by saying that $P(x)$ is true for *almost every* x in S . If $S = X$ we say briefly 'a.e.' instead of 'a.e. in X '.

4.2.4. *Examples.* Let $f, g : X \rightarrow \overline{\mathbb{R}}$.

(i) ' $f = g$ a.e.' means that the set $\{x : f(x) \neq g(x)\}$ is negligible, that is, there exists a null set E such that the following (equivalent) conditions hold: (a) $x \notin E \Rightarrow f(x) = g(x)$; (b) $f|_{\mathbb{C}E} = g|_{\mathbb{C}E}$; (c) $f \varphi_{\mathbb{C}E} = g \varphi_{\mathbb{C}E}$.

The relation $f = g$ a.e. is an equivalence relation in the set $\mathcal{F}(X, \overline{\mathbb{R}})$ of all functions $X \rightarrow \overline{\mathbb{R}}$. {Reason: The union of two null sets is null.} We write $[f] = \{g : g = f \text{ a.e.}\}$ for the equivalence class of f for this relation, and $F(X, \overline{\mathbb{R}})$ for the set of all equivalence classes.

(ii) ' $f \leq g$ a.e.' means that the set $\{x : f(x) > g(x)\}$ is negligible. Equivalently, there exists a null set E such that $f \varphi_{\mathbb{C}E} \leq g \varphi_{\mathbb{C}E}$.

The relation $f \leq g$ a.e. is a preordering of $\mathcal{F}(X, \overline{\mathbb{R}})$ (cf. 1.7.1). Moreover,

$$f = g \text{ a.e.} \Leftrightarrow f \leq g \text{ a.e. and } g \leq f \text{ a.e.}$$

With notations as in (i), the relation $[f] \leq [g]$ defined by $f \leq g$ a.e. is a partial ordering of $F(X, \overline{\mathbb{R}})$ (cf. 1.7.7).

(iii) ' f is constant a.e.' means that there is a $c \in \overline{\mathbb{R}}$ such that $f(x) = c$ a.e., that is, $f = c1$ a.e. in the sense of (i).

(iv) ' f is bounded a.e.' means that there exists a real number $M \geq 0$ such that $|f| \leq M$ a.e. This is usually expressed by saying that f is *essentially bounded* (with respect to μ). The smallest such number M (Exercise 2) is called the *essential supremum* of $|f|$ and is denoted $\text{ess sup } |f|$.

(v) Suppose f satisfies a collection \mathcal{C} of conditions. We say that f is *essentially unique* (or a.e. unique) with respect to \mathcal{C} in case

$$g \text{ satisfies the conditions of } \mathcal{C} \Rightarrow g = f \text{ a.e.};$$

in other words, the set of all functions satisfying the conditions \mathcal{C} is a subset of the equivalence class $[f]$.

In the next examples, (f_n) is a sequence of functions $X \rightarrow \overline{\mathbb{R}}$.

(vi) ' $f_n \rightarrow f$ a.e.' means that there exists a null set E such that

$$x \in X - E \Rightarrow f_n(x) \rightarrow f(x) \text{ in } \overline{\mathbb{R}}$$

(cf. 1.16.8), in other words, $f_n \varphi_{\mathbb{C}E} \rightarrow f \varphi_{\mathbb{C}E}$ pointwise on X .

(vii) The sequence (f_n) is said to be 'Cauchy a.e.' if, for almost every x in X , $(f_n(x))$ is a Cauchy sequence of real numbers.

4.2.5. Theorem. Let f, g, f_n, g_n ($n = 1, 2, 3, \dots$) be real-valued functions on X .

- (1) If $f_n \rightarrow f$ a.e., then the sequence (f_n) is Cauchy a.e.
- (2) If $f_n \rightarrow f$ a.e. and $f_n \rightarrow g$ a.e., then $f = g$ a.e.
- (3) If $f_n \rightarrow f$ a.e. and $f = g$ a.e., then $f_n \rightarrow g$ a.e.
- (4) If $f_n \rightarrow f$ a.e. and if, for each n , $g_n = f_n$ a.e., then $g_n \rightarrow f$ a.e.

Proof. (1) Obvious.

(2) If E and F are null sets such that $f_n \rightarrow f$ on $\mathbb{C}E$ and $f_n \rightarrow g$ on $\mathbb{C}F$, then $E \cup F$ is a null set on whose complement $f(x) = \lim f_n(x) = g(x)$.

(3) Similar to the proof of (2).

(4) The union of a sequence of null sets is null. \diamond

Item (2) says that a.e. limits of sequences are a.e. unique. Some other useful properties of convergence a.e. are listed in the following theorem:

4.2.6. Theorem. With notations as in 4.2.5, suppose $f_n \rightarrow f$ a.e. and $g_n \rightarrow g$ a.e. Let $c \in \mathbb{R}$ and $A \subset X$. Then:

- (5) $cf_n \rightarrow cf$ a.e.
- (6) $f_n + g_n \rightarrow f + g$ a.e.
- (7) $|f_n| \rightarrow |f|$ a.e.
- (8) $f_n \cup g_n \rightarrow f \cup g$ a.e. and $f_n \cap g_n \rightarrow f \cap g$ a.e.
- (9) $f_n^+ \rightarrow f^+$ a.e. and $f_n^- \rightarrow f^-$ a.e.
- (10) $\varphi_A f_n \rightarrow \varphi_A f$ a.e.
- (11) $f_n g_n \rightarrow fg$ a.e.
- (12) If $f_n \leq g_n$ a.e. for all n , then $f \leq g$ a.e.
- (13) If $|f_n| \leq g$ a.e. for all n , then $|f| \leq g$ a.e.
- (14) If $f_n \leq f_{n+1}$ a.e. for all n , then $f_n \uparrow f$ a.e.

Proof. (5), (7) and (10) are obvious.

(6), (11) The union of two null sets is null.

(8), (9) Immediate from (5)–(7) and the formulas in 4.1.13 and 4.1.14.

(12), (14) The union of a sequence of null sets is null.

(13) Since $|f_n| \rightarrow |f|$ a.e. by (7), this is immediate from (12). \diamond

The next theorem is a kind of 'a.e. Cauchy criterion' for sequences of functions:

4.2.7. Theorem. Suppose $f_n : X \rightarrow \mathbb{R}$ ($n = 1, 2, 3, \dots$).

If (f_n) is Cauchy a.e., then there exists a function $f : X \rightarrow \mathbb{R}$ such that $f_n \rightarrow f$ a.e.; if, moreover, the f_n are measurable, then f can be taken to be measurable.

Proof. Let E be a null set such that the sequence $(f_n(x))$ is Cauchy for every $x \in X - E$, and define $f : X \rightarrow \mathbb{R}$ by the formulas

$$f(x) = \begin{cases} \lim f_n(x) & \text{for } x \in X - E \\ 0 & \text{for } x \in E. \end{cases}$$

Clearly $f_n \rightarrow f$ a.e. and $f_n \varphi_{\mathbb{C}E} \rightarrow f \varphi_{\mathbb{C}E} = f$ (pointwise on X). If the f_n are measurable, then so are the $f_n \varphi_{\mathbb{C}E}$, therefore so is their pointwise limit f by (iii) of 4.1.20. \diamond

Exercises

1. Let (X, \mathcal{S}, μ) be a measure space.

(i) The set \mathcal{S}_0 of null sets is an ideal in the Boolean ring \mathcal{S} of measurable sets (cf. §2.4, Exercise 6). (Since \mathcal{S}_0 is closed under countable unions, it is called a σ -ideal of \mathcal{S} .)

(ii) The set \mathcal{N} of negligible sets is a σ -ideal of the Boolean ring $\mathcal{P}(X)$.

2. With notations as in 4.2.4, (iv), if $f : X \rightarrow \overline{\mathbb{R}}$ is essentially bounded then there exists a smallest real number $M \geq 0$ such that $|f| \leq M$ a.e. (This number is called the *essential supremum* of f , often denoted $\|f\|_\infty$.)

{Hint: Let M_0 be the infimum of all such M and choose a sequence of real numbers $M_n \geq 0$ such that $|f| \leq M_n$ a.e. ($n = 1, 2, 3, \dots$) and $M_n \rightarrow M_0$.}

3. The concept of ‘almost everywhere continuity’ is delicate. Let (X, \mathcal{S}, μ) be a measure space and suppose that X is also a topological space. Let Y be another topological space and let $f : X \rightarrow Y$. We say that f is *continuous a.e.* if its set of discontinuities is negligible, that is, the set

$$\{x \in X : f \text{ is not continuous at } x\}$$

is negligible. This means that there exists a null set E such that f is continuous at every $x \in X - E$.

(i) If there exists a null set E such that the restriction of f to $X - E$ is a continuous function $X - E \rightarrow Y$ for the relative topology on $X - E$ (cf. §3.5, Exercise 1), it does not necessarily follow that f is continuous a.e.

(ii) If there exists a continuous function $g : X \rightarrow Y$ such that $f = g$ a.e., it does not necessarily follow that f is continuous a.e.

{Hint: In the context of Lebesgue measure, let $f : \mathbb{R} \rightarrow \mathbb{R}$ be the characteristic function of the rationals.}

4. Suppose (X, \mathcal{S}, μ) is a *complete* measure space (§2.4, Exercise 8). If $f, g : X \rightarrow \mathbb{R}$ are functions such that f is measurable and $g = f$ a.e., then g is also measurable.

{Hint: Let $N \in \mathcal{S}$ with $\mu(N) = 0$ and $g = f$ on $X - N$. If $c \in \mathbb{R}$ and $A = \{x : f(x) > c\}$, $B = \{x : g(x) > c\}$, then $A - B \subset N$, $B - A \subset N$ and $B = [A - (A - B)] \cup (B - A)$.}

4.3. Integrable Simple Functions

As in the preceding section, (X, \mathcal{S}, μ) is a fixed measure space.

The simplest functions to integrate are the simple functions that vanish outside a set of finite measure. First, a useful notation:

4.3.1. *Definition.* For any function $f : X \rightarrow \mathbb{R}$ we write

$$N(f) = \{x \in X : f(x) \neq 0\},$$

called the **nonzero set** of f .

4.3.2. *Remark.* If f is measurable then $N(f) \in \mathcal{S}$, because $N(f) = f^{-1}(\mathbb{R} - \{0\})$ is the inverse image of a Borel set.

4.3.3. *Definition.* An **integrable simple function** (ISF) is a simple function f whose nonzero set has finite measure, that is, $\mu(N(f)) < +\infty$.

4.3.4. *Theorem.* If f and g are ISF, $a \in \mathbb{R}$ and $E \in \mathcal{S}$, then the following functions are also ISF :

$$f + g, af, |f|, f \cup g, f \cap g, f^+, f^-, fg \text{ and } \varphi_E f.$$

In particular, the ISF form an algebra of functions for the operations of pointwise sum, product and scalar multiple.

Proof. If h is any of the functions in question, then h is simple (4.1.22) and $N(h) \subset N(f) \cup N(g)$, where $N(f) \cup N(g)$ has finite measure. \diamond

4.3.5. *Remark.* Since $N(fg) = N(f) \cap N(g) \subset N(f)$, it is clear that if f is an ISF and g is any simple function, then fg is an ISF; thus, the ISF form an ideal in the algebra of simple functions.

Here are some useful characterizations of the integrable simple functions:

4.3.6. *Theorem.* For a function $f : X \rightarrow \mathbb{R}$, the following conditions are equivalent:

- (a) f is an ISF;
- (b) $f(X)$ is finite and, for each nonzero real number c , the set $\{x : f(x) = c\}$ is a measurable set of finite measure;
- (c) f is a linear combination $f = \sum_{i=1}^n c_i \varphi_{E_i}$, where E_1, \dots, E_n are pairwise disjoint measurable sets of finite measure;
- (d) f is a linear combination of characteristic functions φ_E with E measurable of finite measure.

Proof. (a) \Rightarrow (b): For every $c \in \mathbb{R} - \{0\}$, $\{x : f(x) = c\} = f^{-1}(\{c\}) \subset N(f)$.

(b) \Rightarrow (c): If c_1, \dots, c_n are the distinct nonzero values of f and if $E_i = f^{-1}(\{c_i\})$, then $f = \sum_{i=1}^n c_i \varphi_{E_i}$. (If $f = 0$ then $f = 1 \cdot \varphi_\emptyset$ does the job.)

(c) \Rightarrow (d): Trivial.

(d) \Rightarrow (a): Immediate from Theorem 4.3.4. \diamond

If E is a measurable set of finite measure, then the natural ‘integral’ of φ_E is $\mu(E)$, thus if integration is to be a linear operation then the integral of an ISF f is foreordained: if $f = \sum_{i=1}^n c_i \varphi_{E_i}$ as in the above condition (c), then the sum $\sum c_i \mu(E_i)$ is the natural candidate for the integral of f . The following lemma assures that this sum is independent of the particular ‘disjoint representation’ of f :

4.3.7. Lemma. *If*

$$f = \sum_{i=1}^n c_i \varphi_{E_i} = \sum_{j=1}^m d_j \varphi_{F_j},$$

where the E_i, F_j are measurable sets of finite measure, the E_i are pairwise disjoint, the F_j are pairwise disjoint, and c_i, d_j are real numbers, then

$$\sum_{i=1}^n c_i \mu(E_i) = \sum_{j=1}^m d_j \mu(F_j).$$

Proof. We can suppose that the E_i, F_j are nonempty and the c_i, d_j are nonzero. It is then clear that $N(f) = \bigcup E_i = \bigcup F_j$, and

$$E_i = E_i \cap N(f) = \bigcup_{j=1}^m E_i \cap F_j,$$

$$F_j = F_j \cap N(f) = \bigcup_{i=1}^n E_i \cap F_j.$$

Since the mn sets $G_{ij} = E_i \cap F_j$ are pairwise disjoint, we have

$$\varphi_{E_i} = \sum_{j=1}^m \varphi_{G_{ij}}, \quad \varphi_{F_j} = \sum_{i=1}^n \varphi_{G_{ij}}$$

and

$$\mu(E_i) = \sum_{j=1}^m \mu(G_{ij}), \quad \mu(F_j) = \sum_{i=1}^n \mu(G_{ij}).$$

Then

$$\begin{aligned} f &= \sum_{i=1}^n c_i \varphi_{E_i} = \sum_{i=1}^n c_i \left(\sum_{j=1}^m \varphi_{G_{ij}} \right) = \sum_{i,j=1}^{n,m} c_i \varphi_{G_{ij}}, \\ &= \sum_{j=1}^m d_j \varphi_{F_j} = \sum_{j=1}^m d_j \left(\sum_{i=1}^n \varphi_{G_{ij}} \right) = \sum_{i,j=1}^{n,m} d_j \varphi_{G_{ij}}. \end{aligned}$$

It follows from the pairwise disjointness of the G_{ij} that if $G_{ij} \neq \emptyset$ then

$$c_i = \text{the value of } f \text{ on } G_{ij} = d_j,$$

whereas if $G_{ij} = \emptyset$ then $\mu(G_{ij}) = 0$; either way,

$$c_i \mu(G_{ij}) = d_j \mu(G_{ij}) \quad \text{for all } i, j,$$

therefore

$$\begin{aligned} \sum_{i=1}^n c_i \mu(E_i) &= \sum_{i=1}^n c_i \left(\sum_{j=1}^m \mu(G_{ij}) \right) \\ &= \sum_{i,j} c_i \mu(G_{ij}) \\ &= \sum_{i,j} d_j \mu(G_{ij}) \\ &= \sum_{j=1}^m d_j \left(\sum_{i=1}^n \mu(G_{ij}) \right) \\ &= \sum_{j=1}^m d_j \mu(F_j). \quad \diamond \end{aligned}$$

4.3.8. *Definition.* If f is an ISF and $f = \sum_{i=1}^n c_i \varphi_{E_i}$ as in the lemma, we define the **integral** $I(f)$ of f (with respect to the measure μ) by the formula

$$I(f) = \sum_{i=1}^n c_i \mu(E_i)$$

(well-defined, by the lemma).

4.3.9. *Remark.* If f is an ISF then

$$I(f) = \sum_{c \in f(X)} c \mu(f^{-1}(\{c\}))$$

(if $f^{-1}(\{0\})$ has infinite measure, then the convention $0 \cdot (+\infty) = 0$ saves the day). This formula could have been used to define I , but the proof of the additivity of I will require Lemma 4.3.7 anyway.

4.3.10. *Theorem.* The correspondence $f \mapsto I(f)$ is a positive linear form on the vector space of integrable simple functions, that is,

- (1) $I(f + g) = I(f) + I(g)$,
- (2) $I(cf) = cI(f)$, and
- (3) $f \geq 0 \Rightarrow I(f) \geq 0$.

Proof. Let f and g be ISF, $c \in \mathbb{R}$. Each of the functions f and g has a disjoint representation as in (c) of 4.3.6, say

$$f = \sum_{i=1}^n c_i \varphi_{E_i}, \quad g = \sum_{j=1}^m d_j \varphi_{F_j}.$$

Let G_1, \dots, G_r be a listing of the $r = mn$ pairwise disjoint sets $E_i \cap F_j$. By the argument in 4.3.7,

$$f = \sum_{k=1}^r a_k \varphi_{G_k}, \quad g = \sum_{k=1}^r b_k \varphi_{G_k},$$

where each a_k is some c_i , and each b_k is some d_j ; we now have, so to speak, simultaneous disjoint representations of f and g . Also

$$f + g = \sum_{k=1}^r (a_k + b_k) \varphi_{G_k}, \quad cf = \sum_{k=1}^r ca_k \varphi_{G_k}.$$

By Definition 4.3.8,

$$\begin{aligned} I(f + g) &= \sum_{k=1}^r (a_k + b_k) \mu(G_k) \\ &= \sum_{k=1}^r a_k \mu(G_k) + \sum_{k=1}^r b_k \mu(G_k) = I(f) + I(g), \end{aligned}$$

and similarly $I(cf) = cI(f)$.

If $f \geq 0$ then for each i , $c_i \varphi_{E_i} = \varphi_{E_i} f \geq 0$, therefore either $c_i \geq 0$ or $E_i = \emptyset$; in either case, $c_i \mu(E_i) \geq 0$, whence $I(f) \geq 0$. \diamond

4.3.11. Corollary. $|I(f)| \leq I(|f|)$ for every ISF f .

Proof. By the theorem, $f \leq g \Rightarrow I(f) \leq I(g)$, thus the corollary follows on applying I to the inequalities $-|f| \leq f \leq |f|$. \diamond

4.3.12. Corollary. If f is an ISF and M is a constant such that $|f(x)| \leq M < +\infty$ for all x , then $I(|f|) \leq M\mu(N(f))$.

Proof. Apply I to the inequality $|f| \leq M\varphi_{N(f)}$. \diamond

4.3.13. Corollary. The following conditions on a function $f : X \rightarrow \mathbb{R}$ are equivalent:

- (a) f is simple and $f = 0$ a.e.;
- (b) f is an ISF and $I(|f|) = 0$.

Proof. (a) \Rightarrow (b): By assumption $N(f)$ is a null set, therefore f is an ISF and $I(|f|) = 0$ by the preceding corollary.

(b) \Rightarrow (a): Let c be any nonzero value of f and let $E = f^{-1}(\{c\})$. Since $f\varphi_E = c\varphi_E$, we have $|c|\varphi_E = |f|\varphi_E \leq |f|$, therefore $|c|\mu(E) \leq I(|f|) = 0$, whence $\mu(E) = 0$. Since f has only finitely many values, this shows that $N(f)$ is a finite union of null sets. \diamond

4.3.14. Corollary. Let f and g be integrable simple functions. Then:

- (1) $f = g$ a.e. $\Rightarrow I(f) = I(g)$.
- (2) $f \geq 0$ a.e. $\Rightarrow I(f) \geq 0$.
- (3) $f \leq g$ a.e. $\Rightarrow I(f) \leq I(g)$.
- (4) If $f \geq 0$ a.e. and $I(f) = 0$, then $f = 0$ a.e.

Proof. (1) If $f = g$ a.e. then $f - g = 0$ a.e., therefore $I(|f - g|) = 0$ by the preceding corollary, and $I(f - g) = 0$ by 4.3.11.

(2) If E is a null set such that $f\varphi_E \geq 0$ then $f = f\varphi_E$ a.e. and, by (1), $I(f) = I(f\varphi_E) \geq 0$.

(3) Immediate from (2).

(4) By assumption, $|f| = f$ a.e.; by (1), $I(|f|) = I(f) = 0$, thus $f = 0$ a.e. by the preceding corollary. \diamond

The key property of the positive linear form I needed for integrating more general functions is its 'order continuity' in the following sense:

4.3.15. Theorem. If f and f_n ($n = 1, 2, 3, \dots$) are integrable simple functions and if $f_n \uparrow f$ pointwise, then $I(f_n) \uparrow I(f)$.

Proof. Writing $g_n = f - f_n$, we have a sequence (g_n) of ISF such that $g_n \downarrow 0$ pointwise, and our problem is to show that $I(g_n) \downarrow 0$. At any rate, $(I(g_n))$ is a decreasing sequence with $I(g_n) \geq 0$ for all n ; we need only show that $\inf I(g_n) = 0$.

Let $F = N(g_1)$ and let M be an upper bound for g_1 . For every positive integer n , $g_n \leq g_1 \leq M\varphi_F$. Let $\alpha = \inf\{I(g_n) : n = 1, 2, 3, \dots\}$; given any $\epsilon > 0$, it will suffice to show that $\alpha \leq \epsilon\mu(F)$. Let

$$E_n = \{x : g_n(x) \geq \epsilon\};$$

since $g_n \downarrow 0$ pointwise, it follows that $E_n \downarrow \emptyset$. Also $E_n \subset E_1 \subset N(g_1) = F$, where F has finite measure, therefore $\mu(E_n) \rightarrow 0$ (2.6.3). Since $E_n \subset F$ and $g_n(x) < \epsilon$ on the complement of E_n , we have

$$\begin{aligned} g_n &= g_n\varphi_F = g_n\varphi_{F-E_n} + g_n\varphi_{E_n} \\ &\leq \epsilon\varphi_{F-E_n} + g_1\varphi_{E_n} \leq \epsilon\varphi_F + M\varphi_{E_n}, \end{aligned}$$

therefore

$$\alpha \leq I(g_n) \leq \epsilon\mu(F) + M\mu(E_n),$$

and passage to the limit yields $\alpha \leq \epsilon\mu(F)$. \diamond

4.4. Integrable Functions

As in the preceding section, (X, \mathcal{S}, μ) is a fixed measure space and $f \mapsto I(f)$ is the positive linear form defined by μ on the vector space of integrable simple functions (4.3.10).

Our goal in this section is to *extend* I to a (in general) much larger and more useful vector space of functions, to be called ‘integrable’; to avoid any appearance of circularity in the use of the term ‘integrable’, we shall refer to the ‘integrable simple functions’ of the preceding section as ISF’s (any possible ambiguity is dispelled in Corollary 4.4.15 below).

It is technically convenient to begin with positive functions:

4.4.1. *Definition.* A function $f : X \rightarrow \mathbb{R}$ such that $f \geq 0$ on X is said to be **integrable** (with respect to the measure μ) if it is the pointwise limit of an increasing sequence (f_n) of nonnegative ISF’s such that $I(f_n)$ is bounded, that is,

$$0 \leq f_n \uparrow f \text{ pointwise on } X \text{ and } \sup_n I(f_n) < +\infty.$$

Such a function f is measurable (4.1.20). In view of the ‘order continuity’ of I (4.3.15) it would be natural to define the ‘integral’ of f to be the supremum of the $I(f_n)$; this is legitimized by the following lemma:

4.4.2. *Lemma.* With f and (f_n) as in the preceding definition, suppose also that (g_n) is a sequence of ISF’s such that $0 \leq g_n \uparrow f$. Then the sequence $I(g_n)$ is also bounded and

$$\sup_n I(g_n) = \sup_n I(f_n).$$

Proof. If m is a fixed positive integer, then

$$f_n \cap g_m \uparrow f \cap g_m = g_m \text{ as } n \uparrow \infty,$$

therefore (by 4.3.15)

$$(i) \quad I(f_n \cap g_m) \uparrow I(g_m) \text{ as } n \uparrow \infty.$$

Similarly, for each fixed n ,

$$(ii) \quad I(f_n \cap g_m) \uparrow I(f_n) \text{ as } m \uparrow \infty.$$

By the ‘associativity of sups’ (cf. §1.15, Exercise 7),

$$\sup_m \left(\sup_n I(f_n \cap g_m) \right) = \sup_n \left(\sup_m I(f_n \cap g_m) \right),$$

that is, in view of (i) and (ii),

$$\sup_m I(g_m) = \sup_n I(f_n) < +\infty.$$

The details are as follows. Let $a = \sup_n I(f_n)$. Since $I(f_n \cap g_m) \leq I(f_n) \leq a$ for all n and m , by (i) we have

$$I(g_m) \leq a \text{ for all } m,$$

so that the sequence $I(g_m)$ is bounded and

$$\sup I(g_m) \leq a = \sup I(f_n).$$

The sequences (f_n) and (g_m) are now on an equal footing, whence the reverse inequality $\sup I(f_n) \leq \sup I(g_m)$. \diamond

4.4.3. *Definition.* With notations as in 4.4.1, we define (as authorized by the preceding lemma)

$$\int f d\mu = \sup_n I(f_n).$$

Since I is a positive linear form (4.3.10), $I(f_n) \uparrow \int f d\mu$ and $\int f d\mu \geq 0$.

4.4.4. *Remark.* If f is an ISF and $f \geq 0$, then f is integrable in the sense of 4.4.1 and $\int f d\mu = I(f)$.

{Proof: Consider the sequence $f_n = f$ for all n .}

The functional $f \mapsto \int f d\mu$ defined in 4.4.3 is order-preserving in the following strong sense:

4.4.5. *Theorem.* If $0 \leq f \leq g$ with g integrable and f measurable, then f is also integrable and $\int f d\mu \leq \int g d\mu$.

Proof. Let (g_n) be a sequence of ISF's with $0 \leq g_n \uparrow g$ and $I(g_n)$ bounded (4.4.1), and let (f_n) be a sequence of simple functions with $0 \leq f_n \uparrow f$ (4.1.26). Then $(f_n \cap g_n)$ is a sequence of simple functions such that

$$0 \leq f_n \cap g_n \uparrow f \cap g = f.$$

For each n , $0 \leq f_n \cap g_n \leq g_n$ implies that the set $N(f_n \cap g_n) \subset N(g_n)$ has finite measure; thus, the $f_n \cap g_n$ are ISF's and

$$I(f_n \cap g_n) \leq I(g_n) \leq \int g d\mu$$

for all n , therefore f is integrable and

$$\int f d\mu = \sup_n I(f_n \cap g_n) \leq \int g d\mu. \diamond$$

The set of all functions $f \geq 0$ that are integrable in the sense of 4.4.1 is closed under sums and positive scalar multiples:

4.4.6. *Theorem.* If $f \geq 0$, $g \geq 0$ are integrable functions, c is a real number ≥ 0 , and E is a measurable set, then

- (1) $f + g$ is integrable and $\int (f + g) d\mu = \int f d\mu + \int g d\mu$;
- (2) cf is integrable and $\int (cf) d\mu = c \int f d\mu$;
- (3) $\varphi_E f$ is integrable.

Proof. Let (f_n) and (g_n) be sequences of ISF's such that $0 \leq f_n \uparrow f$ and $0 \leq g_n \uparrow g$.

- (1) The $f_n + g_n$ are ISF's, $0 \leq f_n + g_n \uparrow f + g$, and the sequence

$$I(f_n + g_n) = I(f_n) + I(g_n)$$

is bounded, therefore $f + g$ is integrable and passage to the limit in the above equation yields $\int (f + g) d\mu = \int f d\mu + \int g d\mu$.

(2) Similarly the cf_n are ISF's, $0 \leq cf_n \uparrow cf$ and the sequence $I(cf_n) = cI(f_n)$ is bounded, with $cI(f_n) \uparrow c \int f d\mu$.

(3) More generally, suppose $h \geq 0$ is measurable and bounded, say $h(x) \leq M < +\infty$ for all x . Then $0 \leq hf \leq Mf$, where hf is measurable (4.1.13) and Mf is integrable by (2), therefore hf is integrable by the preceding theorem. The assertion of (3) is the special case $h = \varphi_E$. \diamond

Property (1) proves to be the key to extending the definition of integral to real-valued (not necessarily positive) functions.

4.4.7. Definition. A real-valued function $f: X \rightarrow \mathbb{R}$ is said to be **integrable** with respect to μ (briefly, ' μ -integrable') if it can be written in the form $f = g - h$, where the functions g and h are ≥ 0 and integrable in the sense of Definition 4.4.1. The set of all such functions f is denoted $\mathcal{L}_{\mathbb{R}}^1(\mu)$, or (in the context of a fixed measure) simply \mathcal{L}^1 . (In the context of the measure space $(\mathbb{R}, \mathcal{M}, \lambda)$ of Example 2.4.13, such a function f is said to be **Lebesgue-integrable**.)

We would like to define the 'integral' of such a function f to be the real number $\int g d\mu - \int h d\mu$; property (1) of 4.4.6 assures the independence of this expression from the particular representation $f = g - h$ of f :

4.4.8. Lemma. If $g_1 - h_1 = g_2 - h_2$, where the functions g_i, h_i are ≥ 0 and integrable in the sense of 4.4.1, then

$$\int g_1 d\mu - \int h_1 d\mu = \int g_2 d\mu - \int h_2 d\mu.$$

Proof. Apply (1) of 4.4.6 to the two sides of the equation $g_1 + h_2 = g_2 + h_1$. \diamond

This lemma legitimizes the following definition:

4.4.9. Definition. With notations as in Definition 4.4.7, the **integral** of f with respect to μ , denoted $\int f d\mu$, is defined by the formula

$$\int f d\mu = \int g d\mu - \int h d\mu.$$

The foregoing definition is consistent with the earlier definition (4.4.3) for positive functions:

4.4.10. Remark. A function $f \geq 0$ is integrable in the sense of 4.4.1 if and only if it is integrable in the sense of 4.4.7; the values of $\int f d\mu$ given by 4.4.3 and 4.4.9 are then identical.

{Proof: "Only if": If $f \geq 0$ is integrable in the sense of 4.4.1, then $f = f - 0$ shows that it is integrable in the sense of 4.4.7, and the value of $\int f d\mu$ given in 4.4.9 by the choices $g = f$ and $h = 0$ coincides with its value as defined in 4.4.3.

"If": Conversely, if $f \geq 0$ is integrable in the sense of 4.4.7, say $f = g - h$ as in 4.4.7, then $0 \leq f \leq g + h$; since $g + h$ is integrable in the sense of 4.4.1 (by 4.4.6), so is f (by 4.4.5).}

The set \mathcal{L}^1 of μ -integrable functions is a real vector space and the correspondence $f \mapsto \int f d\mu$ is a positive linear form on \mathcal{L}^1 :

4.4.11. Theorem. If $f, g \in \mathcal{L}^1$ and $c \in \mathbb{R}$ then:

- (1) $f + g \in \mathcal{L}^1$ and $\int (f + g) d\mu = \int f d\mu + \int g d\mu$;
- (2) $cf \in \mathcal{L}^1$ and $\int (cf) d\mu = c \int f d\mu$;
- (3) $f \geq 0 \Rightarrow \int f d\mu \geq 0$.

Proof. Write $f = f_1 - f_2$ and $g = g_1 - g_2$, where the f_i and g_i are ≥ 0 and integrable.

(1) We have $f + g = (f_1 + g_1) - (f_2 + g_2)$; by (1) of 4.4.6, $f_1 + g_1$ and $f_2 + g_2$ are integrable, therefore $f + g$ is integrable in the sense of 4.4.7 and

$$\begin{aligned} \int (f + g) d\mu &= \int (f_1 + g_1) d\mu - \int (f_2 + g_2) d\mu \\ &= \int f_1 d\mu + \int g_1 d\mu - \int f_2 d\mu - \int g_2 d\mu \\ &= \left(\int f_1 d\mu - \int f_2 d\mu \right) + \left(\int g_1 d\mu - \int g_2 d\mu \right) \\ &= \int f d\mu + \int g d\mu. \end{aligned}$$

(2) If $c \geq 0$, consider $cf = cf_1 - cf_2$. If $c \leq 0$ then $cf = (-c)f_2 - (-c)f_1$ is covered by the preceding case.

(3) The implication is clear from 4.4.10. \diamond

4.4.12. Corollary. The following conditions on a measurable function f are equivalent:

- (a) f is integrable;
- (b) $|f|$ is integrable;
- (c) f^+ and f^- are integrable.

For such a function f , $|\int f d\mu| \leq \int |f| d\mu$.

Proof. (a) \Rightarrow (b): Write $f = g - h$ as in Definition 4.4.7. Then $|f|$ is measurable and $|f| \leq g + h$, where $g + h$ is integrable, thus $|f|$ is integrable by 4.4.5. Moreover, $-|f| \leq f \leq |f|$, therefore

$$-\int |f| d\mu \leq \int f d\mu \leq \int |f| d\mu$$

by the preceding theorem, whence $|\int f d\mu| \leq \int |f| d\mu$.

(b) \Rightarrow (c): The nonnegative functions f^+ and f^- are measurable and $|f| = f^+ + f^-$ (4.1.15, 4.1.14), therefore f^+ and f^- are integrable by 4.4.5.

(c) \Rightarrow (a) : Since $f = f^+ - f^-$, f is integrable by the preceding theorem (or by 4.4.10 and 4.4.7). \diamond

4.4.13. Corollary. *If f and g are integrable then so are $f \cup g$ and $f \cap g$.*

Proof. This is immediate from the preceding corollary and the formulas in the proof of 4.1.13. \diamond

4.4.14. Corollary. *If f is an integrable function then, for every real number $\epsilon > 0$, the set*

$$\{x : |f(x)| \geq \epsilon\}$$

is a measurable set of finite measure. Moreover, there exists an increasing sequence (F_n) of measurable sets of finite measure, such that $F_n \uparrow N(f)$ and $f\varphi_{F_n}$ is bounded for every n .

Proof. Replacing f by $|f|$ (4.4.12) we can suppose that $f \geq 0$. In view of 4.4.10, there exists a sequence (f_n) of ISF's such that $0 \leq f_n \uparrow f$ and $I(f_n) \uparrow \int f d\mu$. Given any $\epsilon > 0$, let

$$E_n = \{x : f_n(x) > \epsilon\}, \quad E = \{x : f(x) > \epsilon\};$$

the sets E_n and E are measurable by 4.1.8. Since $f_n \uparrow f$ pointwise, the sequence (E_n) is increasing with union E , briefly $E_n \uparrow E$, therefore $\mu(E_n) \uparrow \mu(E)$.

Clearly $E_n \subset N(f_n)$, so the sets E_n have finite measure. Moreover,

$$\epsilon \varphi_{E_n} \leq f_n \varphi_{E_n} \leq f_n,$$

thus $\epsilon \varphi_{E_n}$ is an ISF and $\epsilon \mu(E_n) \leq I(f_n) \leq \int f d\mu$; passage to the limit in the inequality $\mu(E_n) \leq (1/\epsilon) \int f d\mu$ yields $\mu(E) \leq (1/\epsilon) \int f d\mu < +\infty$. The first assertion of the corollary now follows from the inclusion

$$\{x : |f(x)| \geq \epsilon\} \subset \{x : |f(x)| > \epsilon/2\}.$$

The final assertion of the corollary is verified by the sequence $F_n = \{x : 1/n \leq |f(x)| \leq n\}$. \diamond

The next corollary dispels any ambiguity about the phrase 'integrable simple function':

4.4.15. Corollary. *If f is a simple function, then the following conditions are equivalent:*

- (a) f is an ISF in the sense of 4.3.3;
- (b) f is integrable in the sense of 4.4.7.

For such a function f , $\int f d\mu = I(f)$.

Proof. (a) \Rightarrow (b): Writing $f = f^+ - f^-$ (cf. 4.3.4), it follows from Remark 4.4.4 that f is integrable in the sense of 4.4.7 and that

$$\int f^+ d\mu - \int f^- d\mu = I(f^+) - I(f^-),$$

in other words $\int f d\mu = I(f)$.

(b) \Rightarrow (a): The problem is to show that $N(f)$ has finite measure. Assuming $N(f)$ nonempty, that is, f not identically zero, let $c > 0$ be the smallest nonzero value of $|f(x)|$; then $N(f) \subset \{x : |f(x)| > c/2\}$, therefore $N(f)$ has finite measure by the preceding corollary. \diamond

4.4.16. Corollary. *If f is integrable and g is a measurable function such that $g = f$ a.e., then g is integrable and $\int g d\mu = \int f d\mu$.*

Proof. The function $h = |g - f|$ is measurable, $h \geq 0$ (everywhere) and $h = 0$ a.e.; let us show that h is integrable with integral 0. Let (h_n) be a sequence of simple functions such that $0 \leq h_n \uparrow h$; then $h_n = 0$ a.e., therefore h_n is an ISF with $I(h_n) = 0$ (4.3.13). Thus h is integrable and

$$\int h d\mu = \sup_n I(h_n) = 0.$$

Since $|g - f| = h$ is integrable, it follows from 4.4.12 that $g - f$ is integrable, therefore so is $g = (g - f) + f$; moreover,

$$\left| \int (g - f) d\mu \right| \leq \int |g - f| d\mu = \int h d\mu = 0,$$

whence $\int g d\mu = \int f d\mu$. \diamond

4.4.17. Corollary. *If f is integrable and if $g : X \rightarrow \mathbb{R}$ is measurable and essentially bounded, then gf is integrable; moreover, if $0 \leq M < +\infty$ and $|g| \leq M$ a.e., then*

$$\int |gf| d\mu \leq M \int |f| d\mu.$$

Proof. Let $E \in \mathcal{S}$ be a null set such that $|g| \leq M$ on the complement E' of E . Then

$$|gf| \varphi_{E'} \leq M |f| \varphi_{E'} \leq M |f|,$$

where the left member is measurable and the right member is integrable (4.4.12), therefore $|gf| \varphi_{E'}$ is integrable (4.4.5) and

$$\int |gf| \varphi_{E'} d\mu \leq M \int |f| d\mu.$$

Since $|gf| = |gf| \varphi_{E'}$ a.e., it follows from the preceding corollary that $|gf|$

is integrable and that

$$\int |gf| d\mu = \int |gf| \varphi_{E'} d\mu \leq M \int |f| d\mu.$$

Finally, gf is integrable by Corollary 4.4.12. \diamond

4.4.18. Corollary. *If f is integrable and E is a measurable set, then $f\varphi_E$ is integrable.*

Proof. Put $g = \varphi_E$ in the preceding corollary. \diamond

4.4.19. Corollary. *If f and g are integrable functions such that $f \leq g$ a.e., then $\int f d\mu \leq \int g d\mu$.*

Proof. Let E be a null set such that $f \leq g$ on E' . Then $f\varphi_{E'} \leq g\varphi_{E'}$ on X , both members of the inequality are integrable by the preceding corollary, and

$$\int f d\mu = \int f\varphi_{E'} d\mu \leq \int g\varphi_{E'} d\mu = \int g d\mu$$

by Corollary 4.4.16 and Theorem 4.4.11. \diamond

Every measurable function that is 'almost everywhere dominated' by an integrable function is integrable:

4.4.20. Corollary. *If f is integrable, g is measurable and $|g| \leq |f|$ a.e., then g is integrable and*

$$\left| \int g d\mu \right| \leq \int |f| d\mu.$$

Proof. Let E be a null set such that $|g|\varphi_{E'} \leq |f|\varphi_{E'}$, that is, $|g\varphi_{E'}| \leq |f\varphi_{E'}|$. Since $f\varphi_{E'}$ is integrable (4.4.18), so are $|f\varphi_{E'}|$, $|g\varphi_{E'}|$ and $g\varphi_{E'}$ (by 4.4.12 and 4.4.5), therefore g is integrable (4.4.16) and

$$\left| \int g d\mu \right| \leq \int |g| d\mu \leq \int |f| d\mu$$

by 4.4.12 and the preceding corollary. \diamond

4.4.21. Corollary. (1) *If f is integrable and $\int |f| d\mu = 0$ then $f = 0$ a.e.*

(2) *If f is measurable and $f = 0$ a.e., then f is integrable and $\int |f| d\mu = 0$.*

(3) *If f is integrable, $f \geq 0$ a.e. and $\int f d\mu = 0$, then $f = 0$ a.e.*

Proof. (1) Let (g_n) be a sequence of ISF's such that $0 \leq g_n \uparrow |f|$. For every n , $0 \leq \int g_n d\mu \leq \int |f| d\mu = 0$, therefore $g_n = 0$ a.e. by 4.3.14; since $g_n \rightarrow |f|$ pointwise, it follows that $|f| = 0$ a.e. (4.2.5).

(2) Also $|f| = 0$ a.e., therefore $|f|$ is integrable with $\int |f| d\mu = \int 0 d\mu = 0$ (4.4.16), and f is integrable by 4.4.12.

(3) Since $|f|$ is integrable and $f = |f|$ a.e., it follows from 4.4.16 that $\int |f| d\mu = \int f d\mu = 0$, therefore $f = 0$ a.e. by part (1). \diamond

The next definition abstracts the idea of 'indefinite integral' as a real-valued function on the set of measurable sets:

4.4.22. *Definition.* If f is an integrable function and E is a measurable set, the **integral of f over E** , denoted $\int_E f d\mu$, is defined by the formula

$$\int_E f d\mu = \int \varphi_E f d\mu$$

(the integral on the right exists by 4.4.18).

In the case of Lebesgue measure λ on \mathbb{R} , if $E = [a, b]$ we also write

$$\int_a^b f d\lambda \quad \text{instead of} \quad \int_{[a,b]} f d\lambda,$$

and, for fixed $a \in \mathbb{R}$, the function $F(x) = \int_a^x f d\lambda$ is called an 'indefinite integral' of f . A highlight of the Lebesgue theory is that f can be recovered from its indefinite integral: for almost every x , F is differentiable at x with $F'(x) = f(x)$ (proved in §9 of the next chapter).

For a general measure μ we shall see, via the following lemma, that an integrable function is 'almost characterized' by its indefinite integral.

4.4.23. *Lemma.* If f is integrable, E is a measurable set, $f(x) > 0$ a.e. on E and $\int_E f d\mu = 0$, then $\mu(E) = 0$.

Proof. Since $f\varphi_E$ is integrable, $f\varphi_E \geq 0$ a.e. and $\int f\varphi_E d\mu = 0$, we have $f\varphi_E = 0$ a.e. by 4.4.21. Let F be a null set such that $f(x) > 0$ on $E - F$ and let G be a null set such that $(f\varphi_E)\varphi_{G'} = 0$, that is, $f(x) = 0$ on $E - G$; then $F \cup G$ is a null set and $E - F$, $E - G$ are disjoint (f has different values on them), thus

$$E - (F \cup G) = (E - F) \cap (E - G) = \emptyset.$$

This shows that $E \subset F \cup G$, so that E is also a null set. \diamond

4.4.24. *Theorem.* If f and g are integrable functions such that

$$\int_E f d\mu = \int_E g d\mu \quad \text{for all } E \in \mathcal{S},$$

then $f = g$ a.e.

Proof. Writing $h = f - g$, we have

$$\int_E h d\mu = 0 \quad \text{for all } E \in \mathcal{S}.$$

In particular, if $E = \{x : h(x) > 0\}$ then it follows from the lemma that $\mu(E) = 0$, that is, $h \leq 0$ a.e., in other words $f \leq g$ a.e. By symmetry, $g \leq f$ a.e. as well, therefore $f = g$ a.e. \diamond

Exercises

1. If $f \geq 0$ is measurable and $g \geq 0$ is integrable, then $f \cap g$ is integrable.

2. If f and g are integrable functions such that $\int_E f d\mu = \int_E g d\mu$ for all measurable sets E of finite measure, then $f = g$ a.e.

3. The relation $f = g$ a.e. is an equivalence relation in $\mathcal{L}^1 = \mathcal{L}^1_{\mathbb{R}}(\mu)$. The set $L^1 = L^1_{\mathbb{R}}(\mu)$ of equivalence classes $[f]$ is a vector space in a natural way, and $[f] \mapsto \int f d\mu$ defines a linear form on L^1 (positive in the appropriate sense, where $[f] \geq 0$ signifies that $f \geq 0$ a.e.).

4.5. Monotone Convergence Theorem, Fatou's Lemma

In the preceding section, the extension of the integral from the vector space \mathcal{L} of ISF's to the vector space \mathcal{L}^1 of integrable functions was accomplished via approximation by monotone sequences of ISF's (Definition 4.4.1). The core result of the present section is that nothing new is obtained by repeating the process starting from \mathcal{L}^1 ; so to speak, the passage from \mathcal{L} to \mathcal{L}^1 is a 'completion' process, analogous to the passage from the field \mathbb{Q} of rational numbers to the field \mathbb{R} of real numbers (cf. Theorem 1.8.26). Three highly useful corollaries (the *monotone convergence theorem*, *dominated convergence theorem* and *Fatou's lemma*) are readily derived from the core result, which is as follows:

4.5.1. Lemma. Let $f : X \rightarrow \mathbb{R}$ and suppose there exists a sequence (f_n) of integrable functions such that $f_n \uparrow f$ (pointwise on X) and such that the sequence of integrals $\int f_n d\mu$ is bounded. Then f is integrable and $\int f_n d\mu \uparrow \int f d\mu$.

Proof. At any rate, f is measurable (4.1.20). Replacing f_n by $f_n - f_1$ for all n (and f by $f - f_1$), we can suppose that $0 \leq f_n \uparrow f$. For each n let (g_{nj}) be a sequence of ISF's such that

$$0 \leq g_{nj} \uparrow f_n \quad \text{as } j \uparrow \infty.$$

We thus have a tableau of functions

$$\begin{array}{cccccc} & & & & f \\ & & & & \vdots \\ \dots & & & & \\ g_{31} & g_{32} & g_{33} & \dots & f_3 \\ g_{21} & g_{22} & g_{23} & \dots & f_2 \\ g_{11} & g_{12} & g_{13} & \dots & f_1 \end{array}$$

where each row converges pointwise to its right end and the column on the right end converges to its top. For each n , we consider the $n \times n$ southwest

corner of the tableau and we define g_n to be the pointwise supremum of these n^2 functions:

$$g_n = \sup\{g_{ij} : 1 \leq i, j \leq n\}.$$

The g_n are ISF's (4.3.4) and the sequence (g_n) is increasing, and it is clear that $g_n \leq f_n \leq f$ for all n . It follows from the 'associativity of sups' (§1.15, Exercise 7) that

$$(*) \quad \sup_n g_n = \sup_{i,j} g_{ij} = f;$$

for, writing

$$\begin{aligned} I &= \mathbb{P} \times \mathbb{P} = \{(i, j) : i, j = 1, 2, 3, \dots\} \\ J_n &= \{n\} \times \mathbb{P} = \{(n, j) : j = 1, 2, 3, \dots\} \\ K_n &= \{(i, j) : 1 \leq i, j \leq n\}, \end{aligned}$$

we have $I = \bigcup_n J_n = \bigcup_n K_n$, therefore

$$f = \sup_n f_n = \sup_n \left(\sup_{(i,j) \in J_n} g_{ij} \right) = \sup_{(i,j) \in I} g_{ij}$$

and

$$\sup_n g_n = \sup_n \left(\sup_{(i,j) \in K_n} g_{ij} \right) = \sup_{(i,j) \in I} g_{ij}.$$

In particular, it follows from $(*)$ that $0 \leq g_n \uparrow f$; to prove that f is integrable, we need only show that the sequence $I(g_n)$ is bounded. Indeed, if $\int f_n d\mu \leq M < +\infty$ for all n , then

$$\int g_n d\mu \leq \int f_n d\mu \leq M$$

(because $g_n \leq f_n$). We now know that f is integrable, and

$$\int g_n d\mu = I(g_n) \uparrow \int f d\mu$$

by the definition of $\int f d\mu$; moreover,

$$\int g_n d\mu \leq \int f_n d\mu \leq \int f d\mu,$$

therefore $\int f_n d\mu \uparrow \int f d\mu$. \diamond

The hypotheses of the lemma can be weakened considerably:

4.5.2. Lemma. *Let (f_n) be a sequence of integrable functions such that, for every n , $f_n \leq f_{n+1}$ a.e., and such that the sequence of integrals $\int f_n d\mu$ is bounded.*

Then, for almost every x , the sequence $(f_n(x))$ is increasing and bounded (hence convergent to its supremum).

Proof. For each n , the set

$$E_n = \{x : f_n(x) > f_{n+1}(x)\} = \{x : (f_n - f_{n+1})(x) > 0\}$$

is measurable (4.1.8), and $f_n \leq f_{n+1}$ a.e. implies that E_n is a null set, therefore the set

$$E = \bigcup_{n=1}^{\infty} E_n = \{x : (f_n(x)) \text{ is not increasing}\}$$

is also null. Replacing f_n by $f_n \varphi_{\mathbb{C}E}$, we can suppose that $f_n \uparrow$ (pointwise on X); then replacing f_n by $f_n - f_1$, we can suppose further that $f_n \geq 0$ for all n .

The hypothesis now is that (f_n) is a sequence of integrable functions such that $0 \leq f_n \uparrow$ and such that $\int f_n d\mu$ is bounded. Let

$$\begin{aligned} F &= \{x : (f_n(x)) \text{ is not convergent}\} \\ &= \{x : (f_n(x)) \text{ is not bounded}\} \\ &= \{x : (\forall m \in \mathbb{P}) \exists n \in \mathbb{P} \ni f_n(x) \geq m\} \\ &= \bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \{x : f_n(x) \geq m\}. \end{aligned}$$

Writing

$$F_n(m) = \{x : f_n(x) \geq m\}, \quad F(m) = \bigcup_{n=1}^{\infty} F_n(m),$$

we have $F_n(m) \uparrow F(m)$ for each m , and $F(m) \downarrow F$ as $m \uparrow \infty$. Our problem is to show that $\mu(F) = 0$; it will suffice to show that the $F(m)$ have finite measure and that $\mu(F(m)) \rightarrow 0$ as $m \rightarrow \infty$, for it will then follow from Theorem 2.6.3 that $\mu(F) = \lim_{m \rightarrow \infty} \mu(F(m)) = 0$.

Say $\int f_n d\mu \leq M < +\infty$ for all n . Fix a positive integer m . By the definition of $F_n(m)$,

$$m \varphi_{F_n(m)} \leq f_n \varphi_{F_n(m)} \leq f_n$$

(the second inequality holds because $f_n \geq 0$). The sets $F_n(m)$ have finite measure (4.4.14), thus the above inequalities imply that

$$m \mu(F_n(m)) \leq \int f_n d\mu \leq M$$

for every n ; since $\mu(F_n(m)) \rightarrow \mu(F(m))$ by 2.6.2, we conclude that $m \mu(F(m)) \leq M$. Thus $\mu(F(m)) \leq M/m$ for every m , consequently $\mu(F(m)) \rightarrow 0$ as $m \rightarrow \infty$. \diamond

The essence of these lemmas is conveniently summarized as follows:

4.5.3. Theorem. (**Monotone convergence theorem**) Let (f_n) be a sequence of integrable functions such that $f_n \leq f_{n+1}$ a.e. for every n . The following conditions are equivalent:

- (a) there exists an integrable function f such that $f_n \rightarrow f$ a.e.;
- (b) $\int f_n d\mu$ is bounded.

If the conditions are verified, then $f_n \uparrow f$ a.e. and $\int f_n d\mu \uparrow \int f d\mu$.

Proof. (b) \Rightarrow (a): Replacing f_n by $f_n \varphi_{\mathbb{C}E}$ for a suitable null set E , we can suppose by the preceding lemma that the sequence (f_n) is pointwise increasing and bounded, therefore convergent; defining $f(x) = \lim f_n(x)$ for all $x \in X$, we know from Lemma 4.5.1 that f is integrable and that $\int f_n d\mu \uparrow \int f d\mu$.

(a) \Rightarrow (b): For every n , $f_1 \leq f_n \leq f$ a.e., therefore (4.4.19)

$$\int f_1 d\mu \leq \int f_n d\mu \leq \int f d\mu < +\infty. \diamond$$

4.5.4. Corollary. (**Lebesgue's dominated convergence theorem**) Let g be an integrable function and let (f_n) be a sequence of integrable functions such that $|f_n| \leq g$ a.e. for all n and such that the sequence $(f_n(x))$ is convergent for almost every x .

Then, there exists an integrable function f such that $f_n \rightarrow f$ a.e. Necessarily $\int |f_n - f| d\mu \rightarrow 0$, in particular $\int f_n d\mu \rightarrow \int f d\mu$.

Proof. Multiplying g and the f_n by $\varphi_{\mathbb{C}E}$ for a suitable null set E , we can suppose without loss of generality that $|f_n| \leq g$ (everywhere on X) and that $(f_n(x))$ is convergent for every $x \in X$. Define

$$f(x) = \lim_n f_n(x) \text{ for all } x \in X.$$

Since f is measurable (4.1.20) and $|f| \leq g$, it follows that f is integrable (4.4.20). Then $f_n - f$ is integrable, $|f_n - f| \leq |f_n| + |f| \leq 2g$ and $|f_n - f| \rightarrow 0$ pointwise on X .

The functions $h_n = |f_n - f|$ and $h = 2g$ are integrable, with $0 \leq h_n \leq h$ and $h_n \rightarrow 0$ pointwise on X ; our problem is to show that $\int h_n d\mu \rightarrow 0$. For each n , define

$$H_n = \sup_{i \geq n} h_i.$$

Since $0 \leq H_n \leq h$ and H_n is measurable (4.1.19), it follows that H_n is integrable; moreover, for every $x \in X$,

$$H_n(x) \downarrow \limsup_i h_i(x) = \lim_i h_i(x) = 0$$

(cf. 1.16.7), therefore $\int H_n d\mu \downarrow 0$ by the monotone convergence theorem. Since $0 \leq h_n \leq H_n$, it follows that $\int h_n d\mu \rightarrow 0$. \diamond

The next corollary serves as a lemma to the ‘Riesz-Fischer theorem’ to be proved later on (Chapter 6, §7).

4.5.5. Corollary. (**Fatou’s lemma**) *If (f_n) is a sequence of integrable functions such that $f_n \geq 0$ a.e. for all n , and if*

$$\liminf_n \int f_n d\mu < +\infty,$$

then there exists an integrable function f such that $f = \liminf f_n$ a.e. Moreover,

$$\int f d\mu \leq \liminf_n \int f_n d\mu.$$

Proof. Note that the values of the function $\liminf f_n$ may be infinite (4.1.16). The hypothesis on the sequence of integrals is that $\int f_n d\mu$ does not converge to $+\infty$ in the sense of 1.16.8.

We can suppose without loss of generality that $f_n \geq 0$ (everywhere on X). Define

$$g_n = \inf_{i \geq n} f_i \quad (n = 1, 2, 3, \dots), \quad g = \sup_n g_n = \liminf_i f_i$$

(g may have infinite values); the measurable functions g_n are integrable (because $0 \leq g_n \leq f_n$) and $g_n \uparrow g$ pointwise. For each n ,

$$i \geq n \Rightarrow g_n \leq f_i \Rightarrow \int g_n d\mu \leq \int f_i d\mu,$$

therefore

$$(*) \quad \int g_n d\mu \leq \inf_{i \geq n} \int f_i d\mu \leq \liminf_i \int f_i d\mu < +\infty,$$

thus the (increasing) sequence of integrals $\int g_n d\mu$ is bounded; by the monotone convergence theorem, there exists an integrable function f such that $g_n \uparrow f$ a.e. Already $g_n \uparrow g$ a.e., so $f = g$ a.e., that is, $f = \liminf_n f_n$ a.e. Moreover,

$$\int f d\mu = \sup_n \int g_n d\mu \leq \liminf_n \int f_n d\mu$$

by the monotone convergence theorem and the inequalities (*). \diamond

The following convention is analogous to the notation $\sum_{n=1}^{\infty} a_n = +\infty$ for a divergent positive-term series:

4.5.6. Convention. If f is measurable, $f \geq 0$ a.e. and f is not integrable, we write

$$\int f d\mu = +\infty;$$

this means that if (f_n) is a sequence of simple functions such that $0 \leq f_n \uparrow f$ a.e., then either some (hence every succeeding) f_n fails to be an ISF, or else the f_n are all ISF's but the sequence $\int f_n d\mu$ is unbounded.

Exercise

1. The notational convention in 4.5.6 has the following consequences:

(i) If f and g are measurable functions such that $0 \leq f \leq g$ a.e., then $\int f d\mu \leq \int g d\mu$.

(ii) If f is measurable, then $f \in \mathcal{L}^1 \Leftrightarrow \int |f| d\mu < +\infty$.

(iii) If f and f_n ($n = 1, 2, 3, \dots$) are measurable functions such that $f_n \geq 0$ a.e. and $f_n \uparrow f$ a.e., then $\int f_n d\mu \uparrow \int f d\mu$ (in $\overline{\mathbb{R}}$).

(iv) If f, g are measurable functions ≥ 0 a.e., then $\int (f + g) d\mu = \int f d\mu + \int g d\mu$.

4.6. Monotone Classes

The class $\mathcal{B} = \mathcal{B}(\mathbb{R})$ of Borel subset of \mathbb{R} is the σ -algebra generated by the set of all open intervals in \mathbb{R} (2.4.5); a corollary of the main result of this section is that the restriction to \mathcal{B} of Lebesgue measure is the *only* measure on \mathcal{B} that assigns the measure $b - a$ to every open interval (a, b) . We begin with a useful technique for constructing algebras of sets:

4.6.1. Proposition. *Let X be a set and let \mathcal{C} be a nonempty set of subsets of X with the following properties:*

(1) $A, B \in \mathcal{C} \Rightarrow A \cap B \in \mathcal{C}$;

(2) $A \in \mathcal{C} \Rightarrow A' = X - A$ is the union of a finite number of pairwise disjoint sets in \mathcal{C} .

Then the algebra $\mathcal{A}(\mathcal{C})$ of subsets of X generated by \mathcal{C} coincides with the set \mathcal{A} of all finite disjoint unions of sets in \mathcal{C} .

Proof. It is obvious that $\mathcal{C} \subset \mathcal{A} \subset \mathcal{A}(\mathcal{C})$; to prove that $\mathcal{A}(\mathcal{C}) \subset \mathcal{A}$ it will suffice to show that \mathcal{A} is an algebra of sets. According to Definition 2.4.1, we must show that $\emptyset \in \mathcal{A}$ and that \mathcal{A} is closed under complementation and finite unions.

Suppose $E, F \in \mathcal{A}$, say

$$E = \bigcup_{i=1}^m A_i, \quad F = \bigcup_{j=1}^n B_j,$$

where the A_i (the B_j) are pairwise disjoint sets in \mathcal{C} . Then

$$(a) \quad E \cap F = \bigcup_{i,j} A_i \cap B_j \in \mathcal{A}$$

because the mn sets $A_i \cap B_j$ belong to \mathcal{C} by (1) and are pairwise disjoint; thus \mathcal{A} is closed under finite intersections. Since the A'_i belong to \mathcal{A} by (2), it then follows that

$$(b) \quad E' = \bigcap_{i=1}^m A'_i \in \mathcal{A}.$$

From (a) and (b) we conclude that $\emptyset = E \cap E' \in \mathcal{A}$, and $E \cup F = (E' \cap F')' \in \mathcal{A}$. \diamond

4.6.2. Example. Let $[a, b]$ be a fixed closed interval of \mathbb{R} and let \mathcal{C} be the set of *all* subintervals (not necessarily closed) of $[a, b]$. If $I, J \in \mathcal{C}$ then $I \cap J \in \mathcal{C}$ and $[a, b] - I$ is either a subinterval of $[a, b]$ or the union of two disjoint subintervals. Thus \mathcal{C} satisfies the conditions (1), (2) of the preceding proposition. It follows that the set of all finite disjoint unions of subintervals of $[a, b]$ is the algebra $\mathcal{A}(\mathcal{C})$ of subsets of $[a, b]$ generated by \mathcal{C} .

4.6.3. Example. Let \mathcal{C} and \mathcal{D} be sets of subsets of X and Y , respectively, each satisfying the conditions (1), (2) of 4.6.1, and let

$$\mathcal{P} = \{C \times D : C \in \mathcal{C}, D \in \mathcal{D}\}.$$

The formulas

$$\begin{aligned} (C_1 \times D_1) \cap (C_2 \times D_2) &= (C_1 \cap C_2) \times (D_1 \cap D_2) \\ (X \times Y) - (C \times D) &= [(X - C) \times Y] \cup [C \times (Y - D)] \end{aligned}$$

show that the set \mathcal{P} of subsets of $X \times Y$ satisfies the conditions (1), (2) of 4.6.1. In particular, if \mathcal{C} and \mathcal{D} are algebras of subsets of $X \times Y$, then the set of all finite disjoint unions of ‘rectangles’ $C \times D$ with ‘sides’ $C \in \mathcal{C}$, $D \in \mathcal{D}$ is an algebra of subsets of $X \times Y$ (namely, the algebra generated by the rectangles $C \times D$).

The ‘monotonicity’ referred to in the section heading is a key concept in passing from algebras of sets to σ -algebras; the definition is as follows:

4.6.4. Definition. Let X be a set, $\mathcal{T} \subset \mathcal{P}(X)$ a set of subsets of X . We say that \mathcal{T} is **monotone** (or that \mathcal{T} is a ‘monotone class’) if it is closed under monotone sequential unions and intersections:

$$\left. \begin{array}{l} (E_n) \text{ a sequence in } \mathcal{T} \\ E_n \uparrow E \text{ or } E_n \downarrow E \end{array} \right\} \Rightarrow E \in \mathcal{T}.$$

4.6.5. Example. If μ and ν are two finite measures defined on a σ -algebra \mathcal{S} , then the set

$$\{E \in \mathcal{S} : \mu(E) = \nu(E)\}$$

on which μ and ν agree is a monotone class: it is closed under increasing sequential unions by a general property of measures (2.6.2) and under decreasing sequential intersections by a property of finite measures (2.6.3).

It is obvious that every σ -algebra is monotone. Conversely, every monotone algebra is a σ -algebra (because every countable union is the increasing union of its sequence of finite 'partial unions'). The following result is more subtle:

4.6.6. Theorem. (**Lemma on monotone classes**)¹ If

$$\mathcal{A} \subset \mathcal{T} \subset \mathcal{P}(X),$$

where \mathcal{A} is an algebra of subsets of X and \mathcal{T} is a monotone class, then $S(\mathcal{A}) \subset \mathcal{T}$.

Proof. In words, every monotone class that contains an algebra \mathcal{A} of sets also contains the σ -algebra generated by \mathcal{A} ; this means that $S(\mathcal{A})$ is the *smallest* monotone class containing the algebra \mathcal{A} .

There exist monotone classes containing \mathcal{A} (for example, $S(\mathcal{A})$ or $\mathcal{P}(X)$); let \mathcal{M} be the intersection of them all. Clearly \mathcal{M} is itself a monotone class containing \mathcal{A} , and is the smallest such class. In particular,

$$\mathcal{A} \subset \mathcal{M} \subset \mathcal{T} \text{ and } \mathcal{A} \subset \mathcal{M} \subset S(\mathcal{A}).$$

To prove that $S(\mathcal{A}) \subset \mathcal{T}$, it will suffice to show that $S(\mathcal{A}) = \mathcal{M}$; already $\mathcal{M} \subset S(\mathcal{A})$, so we need only show that $S(\mathcal{A}) \subset \mathcal{M}$. For this, it will suffice to show that \mathcal{M} is an algebra; being monotone, it will then be a σ -algebra containing \mathcal{A} and therefore containing $S(\mathcal{A})$.

We are reduced to proving that \mathcal{M} is closed under finite unions and complementation. To this end, we consider pairs of subsets A, B of X such that

$$(*) \quad A - B, B - A \text{ and } A \cup B \text{ belong to } \mathcal{M}.$$

For example, the condition $(*)$ is satisfied if $A \in \mathcal{A}$ and $B \in \mathcal{A}$. For any set $A \subset X$ we write

$$\mathcal{K}(A) = \{B \subset X : A \text{ and } B \text{ satisfy } (*)\}.$$

Since the condition $(*)$ is symmetric in A and B , we have

$$(i) \quad B \in \mathcal{K}(A) \Leftrightarrow A \in \mathcal{K}(B),$$

and, by the remark following $(*)$,

$$(ii) \quad A \in \mathcal{K}(B) \text{ for all } A, B \in \mathcal{A}.$$

The key to the rest of the proof is the observation that

$$(iii) \quad \text{for every subset } B \text{ of } X, \mathcal{K}(B) \text{ is a monotone class.}$$

¹ P.R. Halmos, *Measure theory* [Van Nostrand, New York, 1950; reprinted Springer-Verlag, New York, 1974], §§4, 5.

For, if (A_n) is a monotone (increasing or decreasing) sequence of sets in $\mathcal{K}(B)$, so that in particular

$$A_n - B, B - A_n, A_n \cup B \in \mathcal{M} \text{ for all } n,$$

then 'passage to the limit' implies that A and B satisfy $(*)$; for example, if $A_n \uparrow A$ then

$$A_n - B \uparrow A - B, B - A_n \downarrow B - A, A_n \cup B \uparrow A \cup B,$$

so that $A - B, B - A, A \cup B$ belong to \mathcal{M} (because \mathcal{M} is a monotone class).

If $B \in \mathcal{A}$ then $\mathcal{A} \subset \mathcal{K}(B)$ by (ii); since $\mathcal{K}(B)$ is monotone, it follows that $\mathcal{M} \subset \mathcal{K}(B)$. Thus,

$$(iv) \quad B \in \mathcal{A} \Rightarrow \mathcal{M} \subset \mathcal{K}(B).$$

If $A \in \mathcal{M}$ and $B \in \mathcal{A}$, then $A \in \mathcal{K}(B)$ by (iv), therefore $B \in \mathcal{K}(A)$ by (i); thus

$$A \in \mathcal{M} \Rightarrow \mathcal{A} \subset \mathcal{K}(A),$$

and, since $\mathcal{K}(A)$ is monotone, we conclude that

$$A \in \mathcal{M} \Rightarrow \mathcal{M} \subset \mathcal{K}(A).$$

In other words,

$$A, B \in \mathcal{M} \Rightarrow A - B, B - A, A \cup B \in \mathcal{M}.$$

In particular, \mathcal{M} is closed under finite unions and (because $X \in \mathcal{A} \subset \mathcal{M}$) complementation. \diamond

4.6.7. Corollary. *Let \mathcal{C} be a nonempty set of subsets of a set X , satisfying the conditions (1), (2) of 4.6.1, and let $\mathcal{S} = \mathcal{S}(\mathcal{C})$ be the σ -algebra generated by \mathcal{C} . If μ and ν are finite measures on \mathcal{S} such that*

$$\mu(A) = \nu(A) \text{ for all } A \in \mathcal{C},$$

then $\mu = \nu$ on \mathcal{S} .

Proof. Let $\mathcal{T} = \{E \in \mathcal{S} : \mu(E) = \nu(E)\}$; our problem is to show that $\mathcal{S} \subset \mathcal{T}$. Let \mathcal{A} be the algebra generated by \mathcal{C} . Obviously $\mathcal{S}(\mathcal{A}) = \mathcal{S}(\mathcal{C}) = \mathcal{S}$, and it is clear from 4.6.1 that $\mathcal{A} \subset \mathcal{T}$; since \mathcal{T} is a monotone class (4.6.5), we conclude from 4.6.6 that $\mathcal{S}(\mathcal{A}) \subset \mathcal{T}$. \diamond

The corollary can be extended to certain not necessarily finite (but not too far from finite!) measures:

4.6.8. Corollary. *Let X be a set and let \mathcal{C} be a set of subsets of X with the following properties:*

$$(1) A, B \in \mathcal{C} \Rightarrow A \cap B \in \mathcal{C};$$

(2') $A, B \in \mathcal{C} \Rightarrow A - B$ is the union of a finite number of pairwise disjoint sets in \mathcal{C} ;

(3) X is the union of a sequence of sets in \mathcal{C} .

Let $\mathcal{S} = \mathcal{S}(\mathcal{C})$ be the σ -algebra generated by \mathcal{C} and suppose that μ and ν are measures on \mathcal{S} such that

$$\mu(A) = \nu(A) < +\infty \text{ for all } A \in \mathcal{C}.$$

Then $\mu = \nu$ on \mathcal{S} .

Proof. Note that the pair of conditions (1), (2') is weaker than (i.e., implied by) the pair (1), (2) of 4.6.1 (because $A - B = A \cap B'$). Note also that if $A, B \in \mathcal{C}$ then it follows from (2') that the set

$$A \cup B = (A - B) \cup B$$

is a finite disjoint union of sets in \mathcal{C} . More generally, if A_1, \dots, A_n is a finite list of sets in \mathcal{C} , then $A_1 \cup \dots \cup A_n$ is a finite disjoint union of sets in \mathcal{C} . The proof is by induction on n : assuming inductively that

$$A \cup \dots \cup A_{n-1} = C_1 \cup \dots \cup C_r$$

with the C_j pairwise disjoint sets in \mathcal{C} , the set

$$\begin{aligned} A_1 \cup \dots \cup A_n &= (C_1 \cup \dots \cup C_r) \cup A_n \\ &= [(C_1 \cup \dots \cup C_r) - A_n] \cup A_n \\ &= \left[\bigcup_{j=1}^r (C_j - A_n) \right] \cup A_n \end{aligned}$$

is a pairwise disjoint union of sets in \mathcal{C} by the condition (2').

By the condition (3), there exists a sequence (A_n) of sets in \mathcal{C} such that $X = \bigcup_{n=1}^{\infty} A_n$. The idea of the rest of the proof is to reduce to the case of finite measures by applying 4.6.7 in each of the sets A_n . Writing $B_n = \bigcup_{k=1}^n A_k$, we have $B_n \cap E \uparrow E$ for every $E \in \mathcal{S}$, so it will suffice to show that $\mu(B_n \cap E) = \nu(B_n \cap E)$ for all n and for all $E \in \mathcal{S}$. Since B_n is known to be a finite disjoint union of sets in \mathcal{C} , we need only show that

$$(*) \quad \mu(A \cap E) = \nu(A \cap E) \text{ for all } A \in \mathcal{C} \text{ and } E \in \mathcal{S}.$$

Fix a set $A \in \mathcal{C}$. The desired relation (*) calls attention to the class of sets

$$A \cap \mathcal{S} = \{A \cap E : E \in \mathcal{S}\},$$

which is easily seen to be a σ -algebra of subsets of A ; moreover, the restrictions of μ and ν to $A \cap \mathcal{S}$ are finite measures. On the other hand, the class

$$A \cap \mathcal{C} = \{A \cap B : B \in \mathcal{C}\}$$

clearly satisfies the conditions (1), (2) of 4.6.1 relative to the set A , and we know that $\mu = \nu$ on $A \cap \mathcal{C}$; in view of 4.6.7, it follows that $\mu = \nu$ on the σ -algebra $\mathcal{S}_A = \mathcal{S}_A(A \cap \mathcal{C})$ of subsets of A generated by the class

$A \cap \mathcal{C}$. To verify the condition (*), which says that $\mu = \nu$ on $A \cap \mathcal{S}$, we need only show that $A \cap \mathcal{S} = \mathcal{S}_A$, that is,

$$(**) \quad A \cap \mathcal{S}(\mathcal{C}) = \mathcal{S}_A(A \cap \mathcal{C}).$$

The argument for this is valid for every every set \mathcal{C} of subsets of X and for every subset $A \subset X$. For, $A \cap \mathcal{S}(\mathcal{C})$ is a σ -algebra on A containing $A \cap \mathcal{C}$, therefore

$$A \cap \mathcal{S}(\mathcal{C}) \supset \mathcal{S}_A(A \cap \mathcal{C});$$

on the other hand, the set $\{E \subset X : A \cap E \in \mathcal{S}_A(A \cap \mathcal{C})\}$ is obviously a σ -algebra on X containing \mathcal{C} , hence also containing $\mathcal{S}(\mathcal{C})$, therefore

$$A \cap \mathcal{S}(\mathcal{C}) \subset \mathcal{S}_A(A \cap \mathcal{C}). \quad \diamond$$

4.6.9. Example. The class \mathcal{C} of all *bounded* intervals (open, closed, semi-closed) in \mathbb{R} clearly satisfies the conditions (1) and (2') of the preceding corollary, and the σ -algebra $\mathcal{S}(\mathcal{C})$ generated by \mathcal{C} is the class $\mathcal{B} = \mathcal{B}(\mathbb{R})$ of Borel sets in \mathbb{R} (2.4.5). The restriction $\mu = \lambda|_{\mathcal{B}}$ of Lebesgue measure λ to the σ -algebra \mathcal{B} is a measure that assigns to each bounded interval its length; by the preceding corollary, μ is the *only* such measure on \mathcal{B} . Indeed, any two measures on \mathcal{B} that are equal and finite on \mathcal{C} must be identical:

4.6.10. Corollary. *If $\mathcal{B} = \mathcal{B}(\mathbb{R})$ is the class of Borel sets in \mathbb{R} , \mathcal{I} is the set of all open intervals in \mathbb{R} , and μ, ν are measures on \mathcal{B} such that*

$$\mu(I) = \nu(I) < +\infty \quad \text{for all } I \in \mathcal{I},$$

then $\mu = \nu$ on \mathcal{B} .

Proof. If J is any closed (or semi-closed) interval in \mathbb{R} , then there exists a sequence (I_n) of open intervals with $I_n \downarrow J$, therefore $\mu(J) = \nu(J)$ by 2.6.3. Thus $\mu = \nu$ on the set \mathcal{C} of all bounded intervals, therefore $\mu = \nu$ on \mathcal{B} by Corollary 4.6.8. \diamond

Exercises

1. Corollary 4.6.10 remains true with “open” replaced by “closed” or by “semiclosed”.

2. Let \mathcal{S} be a σ -algebra of subsets of \mathbb{R} containing the class \mathcal{B} of Borel sets, and let μ be a measure on \mathcal{S} assigning to each open interval its length. Then $\mu = \lambda$ on $\mathcal{S} \cap \mathcal{M}$, where λ is Lebesgue measure and \mathcal{M} is the class of all Lebesgue-measurable sets. In particular, λ is the only measure on \mathcal{M} that assigns to each open interval its length.

{Hint: 4.6.10 and 2.4.15.}

4.7. Indefinite Integrals

The ‘indefinite integrals’ discussed in this section are a generalization to abstract measure spaces of the classical concept (see the remark following 4.4.22). The context is a fixed measure space (X, \mathcal{S}, μ) and, as in 4.4.22, we write

$$\int_E f d\mu = \int \varphi_E f d\mu$$

for every integrable function $f \in \mathcal{L}^1(\mu)$ and every measurable set $E \in \mathcal{S}$. The term ‘a.e.’ is understood to be relative to this measure space (§4.2).

4.7.1. *Definition.* For each $f \in \mathcal{L}^1(\mu)$ we define a real-valued function

$$f \cdot \mu : \mathcal{S} \rightarrow \mathbb{R}$$

on the σ -algebra \mathcal{S} of measurable sets by the formula

$$(f \cdot \mu)(E) = \int_E f d\mu.$$

The ‘set function’ $f \cdot \mu$ is called the **indefinite integral** associated with the μ -integrable function f .

Our goal in this section is to prove some general properties of indefinite integrals and to apply them to the classical special case of Lebesgue measure. In the next section we characterize the set functions $f \cdot \mu$ associated with a *finite* measure μ (the core case of the *Radon-Nikodym theorem*, to be proved in greater generality in Chapter 9). We begin with the simplest formal properties of indefinite integrals:

4.7.2. *Theorem.* Let $f, g \in \mathcal{L}^1(\mu)$ and let $c \in \mathbb{R}$.

(i) $(f + g) \cdot \mu = f \cdot \mu + g \cdot \mu$ and $(cf) \cdot \mu = c(f \cdot \mu)$; thus, the correspondence $f \mapsto f \cdot \mu$ is a linear mapping $\mathcal{L}^1(\mu) \rightarrow \mathcal{F}(\mathcal{S}, \mathbb{R})$.

(ii) $f \cdot \mu = g \cdot \mu \Leftrightarrow f = g$ a.e.

(iii) $f \cdot \mu \geq 0 \Leftrightarrow f \geq 0$ a.e.

(iv) $f \cdot \mu \leq g \cdot \mu \Leftrightarrow f \leq g$ a.e.

(v) $f \cdot \mu = f^+ \cdot \mu - f^- \cdot \mu$, where the set functions $f^+ \cdot \mu, f^- \cdot \mu$ are positive.

(vi) If E and F are measurable sets with $E \cap F = \emptyset$, then $(f \cdot \mu)(E \cup F) = (f \cdot \mu)(E) + (f \cdot \mu)(F)$.

(vii) If $E \in \mathcal{S}$ and $\mu(E) = 0$, then $(f \cdot \mu)(E) = 0$.

Proof. (i) Indefinite integrals $f \cdot \mu$ belong to the vector space $\mathcal{F}(\mathcal{S}, \mathbb{R})$ of real-valued functions $\mathcal{S} \rightarrow \mathbb{R}$, whence the possibility of forming the indicated sums and scalar multiples of them. The formulas in question are immediate consequences of the identities

$$\varphi_E(f + g) = \varphi_E f + \varphi_E g, \quad \varphi_E(cf) = c(\varphi_E f)$$

and the linearity of integration.

(ii) The equivalence is immediate from 4.4.16 and 4.4.24.

(iii), \Leftarrow : The implication is a consequence of 4.4.19.

\Rightarrow : Let $E = \{x : f(x) < 0\}$. By assumption, $\int_E f d\mu = (f \cdot \mu)(E) \geq 0$. However, since $\varphi_E f \leq 0$ by the definition of E , we have also $\int_E f d\mu \leq 0$, therefore $\int_E f d\mu = 0$. Then $\int_E (-f) d\mu = 0$ and $-f > 0$ on E , therefore $\mu(E) = 0$ by 4.4.23, in other words, $f \geq 0$ a.e.

(iv) The equivalence is clear from (i) and (iii).

(v) The equality follows from applying (i) to the decomposition $f = f^+ - f^-$ (4.4.12), and the functions $f^+ \cdot \mu$, $f^- \cdot \mu$ are positive by (iii).

(vi) Since E and F are disjoint, $\varphi_{E \cup F} = \varphi_E + \varphi_F$; the asserted equality then follows from the linearity of integration.

(vii) This follows from the fact that $\varphi_E f = 0$ a.e. (relative to the measure μ). \diamond

It follows from (vi) that the set function $f \cdot \mu$ is 'finitely additive' in the sense of 2.6.1, (1). Better yet:

4.7.3. Theorem. *For every $f \in \mathcal{L}^1(\mu)$, the indefinite integral $f \cdot \mu$ is countably additive in the sense of 2.4.12: if (E_n) is a sequence of pairwise disjoint measurable sets, then*

$$(f \cdot \mu) \left(\bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} (f \cdot \mu)(E_n),$$

and the infinite series on the right converges absolutely. In particular, if $f \geq 0$ a.e., then $f \cdot \mu$ is a (finite) measure on \mathcal{S} .

Proof. Let $F_n = \bigcup_{k=1}^n E_k$, $F = \bigcup_{k=1}^{\infty} E_k$, so that $F_n \uparrow F$.

Suppose first that $f \geq 0$ a.e., so that $f \cdot \mu \geq 0$ by (iii) of the preceding theorem. Then $\varphi_{F_n} f \uparrow \varphi_F f$ a.e., so by the monotone convergence theorem (4.5.3),

$$\int \varphi_{F_n} f d\mu \uparrow \int \varphi_F f d\mu,$$

that is, $(f \cdot \mu)(F_n) \uparrow (f \cdot \mu)(F)$. By (vi) of the preceding theorem, we have

$$(f \cdot \mu)(F_n) = \sum_{k=1}^n (f \cdot \mu)(E_k),$$

thus

$$(f \cdot \mu)(F) = \sup_n (f \cdot \mu)(F_n) = \lim_{n \rightarrow \infty} \sum_{k=1}^n (f \cdot \mu)(E_k);$$

this proves the desired formula and shows that $f \cdot \mu$ is a finite measure (2.4.12).

In the general case, write $f = g - h$ with g and h integrable functions ≥ 0 (for example, $g = f^+$, $h = f^-$). Citing the preceding case, we

have

$$\begin{aligned}
 (f \cdot \mu)(F) &= (g \cdot \mu)(F) - (h \cdot \mu)(F) \\
 &= \sum_{n=1}^{\infty} (g \cdot \mu)(E_n) - \sum_{n=1}^{\infty} (h \cdot \mu)(E_n) \\
 &= \sum_{n=1}^{\infty} [(g \cdot \mu)(E_n) - (h \cdot \mu)(E_n)] = \sum_{n=1}^{\infty} (f \cdot \mu)(E_n).
 \end{aligned}$$

Moreover, the inequality $|(f \cdot \mu)(E_n)| \leq (g \cdot \mu)(E_n) + (h \cdot \mu)(E_n)$ shows that the convergence is absolute. (Alternatively, the series converges for every permutation of its terms—i.e., for every permutation of the E_n —a property that characterizes absolute convergence.) \diamond

4.7.4. Corollary. Let \mathcal{A} be an algebra of sets such that \mathcal{S} is the σ -algebra generated by \mathcal{A} . If f and g are integrable functions such that

$$\int_A f d\mu = \int_A g d\mu \quad \text{for all } A \in \mathcal{A},$$

then $f = g$ a.e.

Proof. Let $h = f - g$. We know that $\int_A h d\mu = 0$ for all $A \in \mathcal{A}$, and the problem is to show that $h = 0$ a.e. Write $h = u - v$ with u, v integrable and ≥ 0 . Then

$$(*) \quad \int_A u d\mu = \int_A v d\mu \quad \text{for all } A \in \mathcal{A},$$

and we are to show that $u = v$ a.e. By the preceding theorem, $u \cdot \mu$ and $v \cdot \mu$ are finite measures on \mathcal{S} , and we know from (*) that $u \cdot \mu = v \cdot \mu$ on \mathcal{A} , therefore $u \cdot \mu = v \cdot \mu$ on $\mathcal{S}(\mathcal{A}) = \mathcal{S}$ by 4.6.7, whence $u = v$ a.e. by (ii) of Theorem 4.7.2. \diamond

In the next corollary, we specialize to the case that (X, \mathcal{S}, μ) is the Lebesgue measure space $(\mathbb{R}, \mathcal{M}, \lambda)$ of 2.4.13:

4.7.5. Corollary. If λ is Lebesgue measure on \mathbb{R} and if f and g are Lebesgue-integrable functions such that

$$\int_a^b f d\lambda = \int_a^b g d\lambda \quad \text{for all } a, b \in \mathbb{R} \quad (a \leq b),$$

then $f = g$ a.e.

Proof. As in the preceding corollary, write $f - g = u - v$, where u, v are ≥ 0 and integrable. By hypothesis, the finite measures $u \cdot \lambda$ and $v \cdot \lambda$ are equal on the set of all closed intervals in \mathbb{R} , and our problem is to show that $u = v$ a.e.

Since singletons are negligible for λ , we see that $u \cdot \lambda$ and $v \cdot \lambda$ are also equal on the set \mathcal{I} of all open intervals; it follows that $u \cdot \lambda = v \cdot \lambda$ on the σ -algebra $\mathcal{B} = \mathcal{S}(\mathcal{I})$ of Borel sets of \mathbb{R} (4.6.10).

If E is any Lebesgue-measurable set, we may write $E = B \cup N$, where B is a Borel set, N is negligible and $B \cap N = \emptyset$ (2.4.17); we know that $(u \cdot \lambda)(B) = (v \cdot \lambda)(B)$ and, by (vii) of 4.7.2, $(u \cdot \lambda)(N) = (v \cdot \lambda)(N) = 0$, therefore $(u \cdot \lambda)(E) = (v \cdot \lambda)(E)$ by the additivity of measures. Then $u = v$ a.e. by (ii) of 4.7.2. \diamond

The following example will be the focus of the next chapter:

4.7.6. *Example.* Let $(\mathbb{R}, \mathcal{M}, \lambda)$ be the Lebesgue measure space (2.4.13) and fix a closed interval $[a, b]$ in \mathbb{R} ($a < b$). The intersections of the sets of \mathcal{M} with $[a, b]$ form a σ -algebra \mathcal{M}_0 of subsets of $[a, b]$,

$$\mathcal{M}_0 = \mathcal{M} \cap [a, b] = \{E \cap [a, b] : E \in \mathcal{M}\}.$$

Moreover, \mathcal{M}_0 is the class of all sets in \mathcal{M} that are contained in $[a, b]$,

$$\mathcal{M}_0 = \{E \in \mathcal{M} : E \subset [a, b]\};$$

this follows from the fact that $[a, b] \in \mathcal{M}$ and \mathcal{M} is closed under finite intersections.

Similarly, if \mathcal{B} is the σ -algebra of Borel sets in \mathbb{R} , then the class

$$\mathcal{B}_0 = \mathcal{B} \cap [a, b] = \{B \cap [a, b] : B \in \mathcal{B}\}$$

is a σ -algebra of subsets of $[a, b]$, and

$$\mathcal{B}_0 = \{B \in \mathcal{B} : B \subset [a, b]\}.$$

If \mathcal{C} is the class of all bounded intervals in \mathbb{R} , then \mathcal{B} is the σ -algebra generated by \mathcal{C} , $\mathcal{B} = \mathcal{S}(\mathcal{C})$ (cf. 4.6.9). Clearly the class $\mathcal{C}_0 = \mathcal{C} \cap [a, b]$ is the class of all subintervals of $[a, b]$ and, by an argument in the proof of 4.6.9, the σ -algebra of subsets of $[a, b]$ generated by \mathcal{C}_0 is \mathcal{B}_0 :

$$\mathcal{S}_{[a,b]}(\mathcal{C} \cap [a, b]) = \mathcal{S}(\mathcal{C}) \cap [a, b] = \mathcal{B} \cap [a, b] = \mathcal{B}_0.$$

The restriction of Lebesgue measure to \mathcal{M}_0 is a finite measure; let us for a moment denote it by λ_0 : $\lambda_0 = \lambda|_{\mathcal{M}_0}$. The finite measure space $([a, b], \mathcal{M}_0, \lambda_0)$ is called *Lebesgue measure on $[a, b]$* , and a function $f : [a, b] \rightarrow \mathbb{R}$ is said to be *Lebesgue-integrable* if $f \in \mathcal{L}^1(\lambda_0)$. To simplify the notation, we shall drop the subscript 0; thus, if f is a Lebesgue-integrable function on $[a, b]$, we shall write simply

$$\int f d\lambda, \quad \text{or} \quad \int_a^b f d\lambda, \quad \text{or} \quad \int_{[a,b]} f d\lambda$$

for the integral $\int f d\lambda_0$, and if $[c, d]$ is a closed subinterval of $[a, b]$ then

the notation

$$\int_c^d f d\lambda$$

replaces the ponderous

$$\int \varphi_{[c,d]} f d\lambda_0,$$

where $\varphi_{[c,d]} : [a, b] \rightarrow \mathbb{R}$ is the characteristic function of $[c, d]$ as a subset of $[a, b]$.

4.7.7. Corollary. If $f : [a, b] \rightarrow \mathbb{R}$ is Lebesgue-integrable and if

$$\int_a^x f d\lambda = 0 \quad \text{for all } x \in [a, b],$$

then $f = 0$ a.e.

Proof. Write $f = u - v$ with u, v Lebesgue-integrable and ≥ 0 on $[a, b]$. Our assumption is that

$$\int_a^x u d\lambda = \int_a^x v d\lambda \quad \text{for all } x \in [a, b]$$

and the problem is to show that $u = v$ a.e. With notations as in the preceding example, consider the measures $u \cdot \lambda$ and $v \cdot \lambda$ on the σ -algebra \mathcal{M}_0 of Lebesgue-measurable subsets of $[a, b]$. By hypothesis, $u \cdot \lambda = v \cdot \lambda$ on the set of intervals $[a, x]$, $a \leq x \leq b$. Both measures obviously vanish on singletons, so from the formula

$$(c, d] = [a, d] - [a, c]$$

we see that $u \cdot \lambda = v \cdot \lambda$ on the set \mathcal{C}_0 of all subintervals of $[a, b]$, and it follows that $u \cdot \lambda = v \cdot \lambda$ on the σ -algebra \mathcal{B}_0 generated by \mathcal{C}_0 (4.6.2, 4.6.7). Finally, $u \cdot \lambda = v \cdot \lambda$ on \mathcal{M}_0 by the argument in Corollary 4.7.5, and $u = v$ a.e. by 4.7.2, (ii). \diamond

Exercises

1. A function $f : [a, b] \rightarrow \mathbb{R}$ is Lebesgue-integrable in the sense of 4.7.6 if and only if the function $g : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$g(x) = \begin{cases} f(x) & \text{for } x \in [a, b] \\ 0 & \text{for } x \in \mathbb{R} - [a, b] \end{cases}$$

is integrable with respect to the Lebesgue measure λ on \mathbb{R} . That is, in the notations of 4.7.6,

$$f \in \mathcal{L}^1(\lambda_0) \Leftrightarrow g \in \mathcal{L}^1(\lambda);$$

and, in this case, $\int f d\lambda_0 = \int g d\lambda$.

2. Let $f : [a, b] \rightarrow \mathbb{R}$ be Lebesgue-integrable.

(i) Show that the function $x \mapsto \int_a^x f d\lambda$ is continuous on $[a, b]$.

{Hint: It suffices to consider $x_n \uparrow x$ and $x_n \downarrow x$.}

(ii) If $\int_a^x f d\lambda = 0$ for all x in a dense subset of $[a, b]$, then $f = 0$ a.e.

(iii) If $\int_a^x f d\lambda = 0$ a.e. then $f = 0$ a.e. and $\int_a^x f d\lambda = 0$ for all x .

{Hint: A negligible set has empty interior; cf. §3.3, Exercise 5.}

3. (i) As in Example 4.7.6, let \mathcal{B}_0 be the set of all Borel subsets of the closed interval $[a, b]$, and let μ be any measure on \mathcal{B}_0 . If f is a μ -integrable function such that

$$\int_a^x f d\mu = 0 \quad \text{for all } x \in [a, b],$$

then $f = 0$ a.e. (relative to μ).

{Hint: Write $f = u - v$ with u, v μ -integrable and ≥ 0 , and let $\mathcal{T} = \{B \in \mathcal{B}_0 : (u \cdot \mu)(B) = (v \cdot \mu)(B)\}$. If $a \leq c \leq d \leq b$, then the formula $(c, d] = [a, d] - [a, c]$ shows that $(c, d] \in \mathcal{T}$. Infer that \mathcal{T} contains every singleton $\{d\}$ ($d \in [a, b]$), hence it contains the set \mathcal{C}_0 of all subintervals of $[a, b]$; conclude that if \mathcal{A} is the algebra of subsets of $[a, b]$ generated by \mathcal{C}_0 , then $\mathcal{A} \subset \mathcal{T}$. Then cite 4.7.4.}

(ii) It suffices to assume in (i) that $\int_a^x f d\mu = 0$ for $x = b$ and for x in a dense subset of (a, b) .

4. If f is a Lebesgue-integrable function on \mathbb{R} , c is a fixed point of \mathbb{R} , and $\int_c^x f d\lambda = 0$ for all x (or for all x in a dense subset of \mathbb{R}), then $f = 0$ a.e.

{Hint: The convention is that when $x < c$, $\int_c^x f d\lambda = -\int_x^c f d\lambda$. Argue that if $a \leq b$ then

$$\int_a^b f d\lambda = \int_c^b f d\lambda - \int_c^a f d\lambda,$$

by considering separately the three cases $c < a \leq b$, $a \leq c \leq b$, $a \leq b < c$.

4.8. Finite Signed Measures

As in the preceding section, (X, \mathcal{S}, μ) is a fixed measure space; the term ‘measurable’ refers to the σ -algebra \mathcal{S} (§4.1) and ‘a.e.’ refers to the measure μ (§4.2). The main result of this section is obtained only for the case that μ is a finite measure (a more general theorem is proved in Chapter 9, §2.)

We saw in the preceding section that every integrable function $f \in \mathcal{L}^1(\mu)$ defines a countably additive, real-valued set function $f \cdot \mu : \mathcal{S} \rightarrow \mathbb{R}$ that vanishes on the null sets for μ . In the present section we shall prove a partial converse: If μ is a *finite* measure, then every countably additive

set function on \mathcal{S} that vanishes on the null sets for μ is of the form $f \cdot \mu$ for a suitable μ -integrable function f . This result plays a strategic role in the differentiation theory exposed in the next chapter (cf. the proof of Theorem 5.2.1).

4.8.1. Definition. A set function $\nu : \mathcal{S} \rightarrow \mathbb{R}$ is called a **finite signed measure** (on \mathcal{S}) if it is countably additive in the following sense: if (E_n) is a sequence of pairwise disjoint sets in \mathcal{S} , then the infinite series $\sum_{n=1}^{\infty} \nu(E_n)$ is convergent and

$$\nu \left(\bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} \nu(E_n).$$

The word ‘finite’ in the definition refers to the fact that ν is real-valued. (In Chapter 9, §1 we shall consider set functions with values in the set $\overline{\mathbb{R}}$ of extended real numbers.)

4.8.2. Examples. (i) If $f \in \mathcal{L}^1(\mu)$ then $f \cdot \mu$ is a finite signed measure on \mathcal{S} (4.7.3).

(ii) If μ_1 and μ_2 are finite measures on \mathcal{S} , then $\mu_1 - \mu_2$ is a finite signed measure on \mathcal{S} . More generally, the set of finite signed measures on \mathcal{S} is closed under real linear combinations, hence is a vector space over \mathbb{R} .

(iii) if ν is a finite signed measure on \mathcal{S} then, for each set $A \in \mathcal{S}$, the set function $E \mapsto \nu(A \cap E)$ is a finite signed measure on \mathcal{S} .

Some useful properties of finite signed measures are listed in the following proposition:

4.8.3. Proposition. Let ν be a finite signed measure on \mathcal{S} .

(i) If (E_n) is a sequence of pairwise disjoint measurable sets, then the series $\sum_{n=1}^{\infty} \nu(E_n)$ converges absolutely.

(ii) $\nu(\emptyset) = 0$.

(iii) ν is finitely additive: if E_1, \dots, E_n are pairwise disjoint measurable sets, then $\nu(\bigcup_{k=1}^n E_k) = \sum_{k=1}^n \nu(E_k)$.

(iv) ν is subtractive: if E and F are measurable sets with $E \subset F$, then $\nu(F - E) = \nu(F) - \nu(E)$.

(v) If (E_n) is a sequence of measurable sets and if $E_n \uparrow E$ or $E_n \downarrow E$, then $\nu(E_n) \rightarrow \nu(E)$.

Proof. (i) See the remark at the end of the proof of 4.7.3 (or Exercise 5 below).

(ii) If $E_n = \emptyset$ for all n , then the constant term series $\sum_{n=1}^{\infty} \nu(E_n)$ is convergent.

(iii) Let $E_k = \emptyset$ for all $k > n$, then cite countable additivity and (ii).

(iv) Since $E = (E - F) \cup F$, we have $\nu(E) = \nu(E - F) + \nu(F)$ by finite additivity; since ν is finite, the term $\nu(F)$ may be transposed. (Matters are otherwise when ν is permitted to have infinite values.)

(v) Suppose, for example, that $E_n \uparrow E$. Writing $E_0 = \emptyset$, we have

$$E = (E_1 - E_0) \cup (E_2 - E_1) \cup (E_3 - E_2) \cup \dots,$$

where the terms of the union on the right side are pairwise disjoint; citing (iv) at the appropriate step,

$$\begin{aligned} \nu(E) &= \sum_{n=1}^{\infty} \nu(E_n - E_{n-1}) \\ &= \lim_{n \rightarrow \infty} \sum_{k=1}^n \nu(E_k - E_{k-1}) \\ &= \lim_{n \rightarrow \infty} \sum_{k=1}^n [\nu(E_k) - \nu(E_{k-1})] \\ &= \lim_{n \rightarrow \infty} [\nu(E_n) - \nu(\emptyset)] = \lim_{n \rightarrow \infty} \nu(E_n). \quad \diamond \end{aligned}$$

It is useful to have a notation for the signed measures described in 4.8.2, (iii):

4.8.4. *Definition.* Let ν be a finite signed measure on \mathcal{S} . For every $A \in \mathcal{S}$ we write ν_A for the finite signed measure on \mathcal{S} defined by the formula $\nu_A(E) = \nu(A \cap E)$ for all $E \in \mathcal{S}$.

4.8.5. *Proposition.* Let ν be a finite signed measure on \mathcal{S} and let $A, B \in \mathcal{S}$.

(i) ν_A is a measure on \mathcal{S} if and only if $\nu_A \geq 0$ (equivalently, $\nu(F) \geq 0$ for every measurable subset F of A).

(ii) $\nu_{\emptyset} = 0$.

(iii) $\nu_{A \cup B} + \nu_{A \cap B} = \nu_A + \nu_B$.

(iv) $A \cap B = \emptyset \Rightarrow \nu_{A \cup B} = \nu_A + \nu_B$.

(v) $A \subset B \Rightarrow \nu_{B-A} = \nu_B - \nu_A$.

(vi) $\nu_{A \cap B} = (\nu_A)_B$.

(vii) When μ is a finite measure, $\mu_A = \varphi_A \cdot \mu$.

Proof. Property (iii) follows from the equality $A \cup B - B = A - A \cap B$ and 4.8.3, (iv). Properties (i), (ii), (vi) are obvious, and (iv) follows from (ii) and (iii). If $A \subset B$ then $B = A \cup (B - A)$, so $\nu_B = \nu_A + \nu_{B-A}$ by (iv), whence (v).

Assuming μ is a finite measure, property (vii) reduces to the computation

$$\mu(A \cap E) = \int \varphi_{A \cap E} d\mu = \int \varphi_A \varphi_E d\mu = \int_E \varphi_A d\mu$$

for all $E \in \mathcal{S}$. \diamond

The vanishing of the indefinite integrals $f \cdot \mu$ on the null sets for μ is codified in the following terminology:

4.8.6. *Definition.* A finite signed measure ν on \mathcal{S} is said to be **absolutely continuous** with respect to the measure μ , written $\nu \ll \mu$, if it vanishes on the null sets for μ , that is,

$$E \in \mathcal{S}, \mu(E) = 0 \Rightarrow \nu(E) = 0.$$

4.8.7. *Examples.* (i) If $f \in \mathcal{L}^1(\mu)$ then $f \cdot \mu \ll \mu$ by 4.7.2, (vii).

(ii) If $\nu \ll \mu$ then $\nu_A \ll \mu$ for every $A \in \mathcal{S}$.

The main result of the section is a partial converse to Example (i): if μ is a *finite* measure then every finite signed measure ν on \mathcal{S} such that $\nu \ll \mu$ is of the form $\nu = f \cdot \mu$ for some μ -integrable function f . A key step in the proof is to decompose a finite signed measure as a difference of two finite measures:

4.8.8. *Theorem. (Hahn decomposition)* If ν is a finite signed measure on the σ -algebra \mathcal{S} of subsets of X , then there exists a set $A \in \mathcal{S}$ such that $\nu_A \geq 0$ and $\nu_{X-A} \leq 0$; thus

$$\nu = \nu_A - (-\nu_{X-A})$$

is the difference of two finite measures.

*Proof.*¹ The idea of the proof is to construct a 'maximal' set $B \in \mathcal{S}$ such that $\nu_B \leq 0$ and take A to be the complement of B .

Let $\mathcal{N} = \{B \in \mathcal{S} : \nu_B \leq 0\}$; at least $\emptyset \in \mathcal{N}$. The proof is organized as a series of remarks about the class \mathcal{N} .

(i) If $B \in \mathcal{N}$ and $E \in \mathcal{S}$ then $B \cap E \in \mathcal{N}$. (Equivalently, if $B \in \mathcal{N}$ then \mathcal{N} contains every measurable subset of B .)

For, since $\nu_B \leq 0$ we have also $\nu_{B \cap E} = (\nu_B)_E \leq 0$.

(ii) If $B, C \in \mathcal{N}$ then $B \cup C \in \mathcal{N}$.

Citing (iii) and (v) of Proposition 4.8.5, we have

$$\nu_{B \cup C} = \nu_B + \nu_C - \nu_{B \cap C} = \nu_B + \nu_{C - B \cap C},$$

where $B \in \mathcal{N}$ and where $C - B \cap C = C \cap (X - B) \in \mathcal{N}$ by (i), thus $\nu_{B \cup C}$ is the sum of two set functions ≤ 0 .

(iii) If $B_n \in \mathcal{N}$ ($n = 1, 2, 3, \dots$) then $\bigcup_{n=1}^{\infty} B_n \in \mathcal{N}$.

Let $B = \bigcup_{n=1}^{\infty} B_n$. Replacing B_n by $B_1 \cup \dots \cup B_n$ (permissible by the preceding remark) we can suppose that $B_n \uparrow B$. For every $E \in \mathcal{S}$ we have $B_n \cap E \uparrow B \cap E$, therefore $\nu(B_n \cap E) \rightarrow \nu(B \cap E)$ by (v) of 4.8.3; since $\nu(B_n \cap E) = \nu_{B_n}(E) \leq 0$ for all n , we conclude that $\nu_B(E) \leq 0$, whence $B \in \mathcal{N}$.

(iv) The set of numbers $\{\nu(B) : B \in \mathcal{N}\}$ has a least element.

¹ Taken from P. R. Halmos' *Measure theory* [Van Nostrand, 1950; reprinted Springer-Verlag, 1974], p. 121, Theorem 29.A. For a shorter proof using Zorn's lemma, see the proof of the decomposition theorem for signed measures (not necessarily finite) in 9.1.16–9.1.20.

Let $\beta = \inf\{\nu(B) : B \in \mathcal{N}\}$. We assert that β is finite, that is, $\beta > -\infty$. It suffices to show that if (B_n) is any sequence in \mathcal{N} then $\nu(B_n)$ is bounded below. Let $B = \bigcup_{n=1}^{\infty} B_n$. We know from (iii) that $-\nu_B$ is a measure; since $B_n \subset B$ for all n , it follows that $-\nu_B(B_n) \leq -\nu_B(B)$, thus the sequence $\nu(B_n)$ is bounded below by $\nu(B)$.

Now let (B_n) be any sequence in \mathcal{N} such that $\nu(B_n) \rightarrow \beta$ and let $B = \bigcup_{n=1}^{\infty} B_n$. By (iii), $B \in \mathcal{N}$; we will show that $\nu(B) = \beta$, which will prove the assertion of (iv). At any rate $\nu(B) \geq \beta$ by the definition of β . On the other hand, $\nu(B_n) \geq \nu(B)$ by the argument in the preceding paragraph, and passage to the limit yields $\beta \geq \nu(B)$, completing the proof of (iv).

For the rest of the proof we fix an element $B \in \mathcal{N}$ such that $\nu(B) = \beta$, that is, $\nu(B) \leq \nu(C)$ for all $C \in \mathcal{N}$.

(v) If $C \in \mathcal{N}$ and $C \subset X - B$ then $\nu(C) = 0$.

For, $B \cup C \in \mathcal{N}$ and, by the minimality of $\nu(B)$,

$$\nu(B) \leq \nu(B \cup C) = \nu(B) + \nu(C) \leq \nu(B),$$

whence equality throughout and $\nu(C) = 0$.

(vi) If $C \in \mathcal{S}$, $C \subset X - B$ and $\nu(C) \neq 0$, then $\nu(E) > 0$ for some measurable set $E \subset C$.

Since the existence of E is equivalent to $C \notin \mathcal{N}$, the assertion (vi) is just (v) in contrapositive form.

For every positive integer n , let

$$\mathcal{E}_n = \{E \in \mathcal{S} : \nu(E) \geq 1/n\}.$$

According to (vi), if $C \in \mathcal{S}$, $C \subset X - B$ and $\nu(C) \neq 0$, then C contains a measurable set E such that $\nu(E) \geq 1/n$ for some positive integer n ; thus, there exists an n such that $\mathcal{P}(C) \cap \mathcal{E}_n \neq \emptyset$, where $\mathcal{P}(C)$ is the class of all subsets of C . Writing

$$\mathbb{P}(C) = \{n \in \mathbb{P} : \mathcal{P}(C) \cap \mathcal{E}_n \neq \emptyset\},$$

we have $\mathbb{P}(C) \neq \emptyset$ and we define $n(C)$ to be the smallest element of $\mathbb{P}(C)$. Summarizing:

(vii) If $C \in \mathcal{S}$, $C \subset X - B$ and $\nu(C) \neq 0$, then there exists a smallest positive integer $n(C)$ such that $\nu(E) \geq 1/n(C)$ for some measurable subset E of C .

To put it another way,

(viii) If $C \in \mathcal{S}$, $C \subset X - B$ and $\nu(C) \neq 0$, then $\nu(E) < 1/n$ for every $n < n(C)$ and every measurable set $E \subset C$.

For, if $n < n(C)$ then $n \notin \mathbb{P}(C)$, that is, $\mathcal{P}(C) \cap \mathcal{E}_n = \emptyset$; thus, for every measurable set $E \subset C$, we have $E \notin \mathcal{E}_n$, in other words $\nu(E) < 1/n$. Expressed in contrapositive form:

(ix) If $C \in \mathcal{S}$, $C \subset X - B$ and $\nu(C) \neq 0$, and if n is a positive integer such that $\nu(E) \geq 1/n$ for some measurable set $E \subset C$, then $n \geq n(C)$.

We have arrived at the conclusive step:

(x) With $B \in \mathcal{N}$ chosen so that $\nu(B) = \beta$, we have $\nu_{X-B} \geq 0$.

If C is any measurable subset of $X-B$, we are to show that $\nu(C) \geq 0$. Assume to the contrary that $\nu(C) < 0$. Since C meets the conditions of (vii), we can define $n_1 = n(C)$ and choose a measurable set $E_1 \subset C$ such that $\nu(E_1) \geq 1/n_1$. Then

$$C - E_1 \subset C \subset X - B$$

and

$$\nu(C - E_1) = \nu(C) - \nu(E_1) < \nu(C) < 0,$$

so $C - E_1$ meets the conditions of (vii); define $n_2 = n(C - E_1)$ and choose a measurable set $E_2 \subset C - E_1$ such that $\nu(E_2) \geq 1/n_2$.

Continuing recursively, we obtain a sequence (E_k) of pairwise disjoint measurable sets such that

$$E_k \subset C - (E_1 \cup \dots \cup E_{k-1})$$

(with the convention $E_0 = \emptyset$) and $\nu(E_k) \geq 1/n_k$, where

$$n_k = n(C - (E_1 \cup \dots \cup E_{k-1})).$$

Let $E = \bigcup_{k=1}^{\infty} E_k$. Then

$$+\infty > \nu(E) = \sum_{k=1}^{\infty} \nu(E_k) \geq \sum_{k=1}^{\infty} 1/n_k,$$

so the series on the right side is convergent and in particular $1/n_k \rightarrow 0$.

Let $F = C - E$. Then $F \subset C \subset X - B$ and

$$\nu(F) = \nu(C) - \nu(E) < \nu(C) < 0,$$

so by (vi) there exists a measurable set $G \subset F$ such that $\nu(G) > 0$. Since $1/n_k \rightarrow 0$, we can choose an index r such that $\nu(G) \geq 1/n_r$ and $n_r > 2$. Since G is disjoint from every E_k , we have

$$G \cup E_r \subset C - (E_1 \cup \dots \cup E_{r-1}) \subset C \subset X - B$$

and

$$\nu(G \cup E_r) = \nu(G) + \nu(E_r) \geq \frac{1}{n_r} + \frac{1}{n_r} > \frac{1}{n_r - 1}$$

(the last inequality reduces to $n_r > 2$), and remark (ix) yields the absurdity

$$n_r - 1 \geq n(C - (E_1 \cup \dots \cup E_{r-1})) = n_r.$$

The assumption $\nu(C) < 0$ has led to a contradiction, therefore $\nu(C) \geq 0$ and the proof that $\nu_{X-B} \geq 0$ is complete; the set $A = X - B$ meets the requirements of the theorem. \diamond

The following two consequences of the Hahn decomposition lead up to the Radon-Nikodym theorem.

4.8.9. Lemma. *If ν is a finite signed measure on \mathcal{S} and if E is a measurable set with $\nu(E) > 0$, then there exists a measurable set $F \subset E$ such that $\nu_F \geq 0$ and $\nu(F) > 0$.*

Proof. With A as in the statement of the preceding theorem,

$$0 < \nu(E) = \nu(A \cap E) + \nu(A' \cap E) \leq \nu(A \cap E)$$

(the last inequality because $\nu_{A'} \leq 0$) and $\nu_{A \cap E} = (\nu_A)_E \geq 0$, so the set $F = A \cap E$ meets the requirements. \diamond

4.8.10. Lemma. *If ν and μ are finite measures on the σ -algebra \mathcal{S} such that $\nu \ll \mu$ and $\nu \neq 0$, then there exists a μ -integrable function $f \geq 0$ such that $f \cdot \mu \leq \nu$ and $f \cdot \mu \neq 0$.*

Proof. By assumption, $0 < \nu(X) < +\infty$; since $\nu \ll \mu$, it follows that $\mu(X) > 0$. Choose $\epsilon > 0$ sufficiently small that $\epsilon\mu(X) < \nu(X)$; then $\rho = \nu - \epsilon\mu$ is a finite signed measure on \mathcal{S} such that $\rho(X) > 0$, so by the preceding lemma there exists a measurable set F such that $\rho_F \geq 0$ and $\rho(F) > 0$. Thus

$$\epsilon\mu_F \leq \nu_F \leq \nu \quad \text{and} \quad \nu(F) - \epsilon\mu(F) > 0.$$

Since $\nu \ll \mu$, it follows that $\mu(F) > 0$; moreover, $\epsilon\mu_F = f \cdot \mu$ with $f = \epsilon\varphi_F$, and $(f \cdot \mu)(F) = \epsilon\mu(F) > 0$, thus $f \cdot \mu \leq \nu$ and $f \cdot \mu \neq 0$. \diamond

4.8.11. Theorem. (**Radon-Nikodym**) *Let μ and ν be finite measures on the σ -algebra \mathcal{S} of subsets of X . In order that $\nu \ll \mu$, it is necessary and sufficient that there exists a μ -integrable function $f \geq 0$ such that $\nu = f \cdot \mu$.*

Proof. We remark that in Chapter 9, §2 the theorem will be generalized to σ -finite measures μ and to signed measures ν that are not necessarily finite (in which case the function f will be measurable but not necessarily μ -integrable). The special case proved here recovers a classical result of Lebesgue in the differentiation theory of the next chapter (Theorem 5.2.1).

Sufficiency. It was noted in Example 4.8.7, (i) that $f \cdot \mu \ll \mu$ (for this, μ need not be finite and f need not be ≥ 0).

Necessity. Assuming $\nu \ll \mu$, we seek a μ -integrable function $f \geq 0$ such that $\nu = f \cdot \mu$. The idea of the proof is to ‘exhaust’ ν by the measures $f \cdot \mu$ that it majorizes (via the preceding lemma). Let

$$\mathcal{K} = \{f \in \mathcal{L}^1(\mu) : f \geq 0 \text{ and } f \cdot \mu \leq \nu\};$$

at least $0 \in \mathcal{K}$. The proof is organized as a series of remarks about \mathcal{K} (the parallelism with the Hahn decomposition theorem will be evident).

(i) *If $f, g \in \mathcal{K}$ then $f \cup g \in \mathcal{K}$.*

Let $h = f \cup g$; thus $h(x) = \max\{f(x), g(x)\}$ for all x , and h is μ -integrable (4.4.13). Given any $E \in \mathcal{S}$, we have to show that $(h \cdot \mu)(E) \leq \nu(E)$. Let

$$F = \{x \in E : h(x) = f(x)\} = E \cap (h - f)^{-1}(\{0\})$$

and let $G = E - F$. Clearly $h(x) = g(x)$ for all $x \in G$, thus

$$\varphi_F h = \varphi_F f \quad \text{and} \quad \varphi_G h = \varphi_G g.$$

Since $\varphi_E = \varphi_F + \varphi_G$, we have

$$\begin{aligned} (h \cdot \mu)(E) &= \int \varphi_E h d\mu = \int \varphi_F h d\mu + \int \varphi_G h d\mu \\ &= \int \varphi_F f d\mu + \int \varphi_G g d\mu \\ &= (f \cdot \mu)(F) + (g \cdot \mu)(G) \leq \nu(F) + \nu(G) = \nu(E). \end{aligned}$$

Thus $h \cdot \mu \leq \nu$, so that $h \in \mathcal{K}$.

(ii) The set $\{\int f d\mu : f \in \mathcal{K}\}$ is bounded.

Indeed, for every $f \in \mathcal{K}$,

$$0 \leq \int f d\mu = (f \cdot \mu)(X) \leq \nu(X) < +\infty.$$

(iii) The set in (ii) has a largest element.

Let $M = \sup\{\int f d\mu : f \in \mathcal{K}\}$; the problem is to show that $M = \int f d\mu$ for some $f \in \mathcal{K}$. Choose a sequence $g_n \in \mathcal{K}$ such that $\int g_n d\mu \rightarrow M$ and let

$$f_n = g_1 \cup \dots \cup g_n \quad (n = 1, 2, 3, \dots).$$

Then $g_n \leq f_n$ and $f_n \in \mathcal{K}$ by (i), therefore

$$\int g_n d\mu \leq \int f_n d\mu \leq M;$$

since the sequence (f_n) is increasing, it follows that $\int f_n d\mu \uparrow M$. By the monotone convergence theorem (4.5.3) there exists a μ -integrable function f such that $f_n \uparrow f$ a.e. (with respect to μ) and we can suppose that $f \geq 0$ everywhere on X . Then

$$\int f d\mu = \sup \int f_n d\mu = M$$

and it remains only to show that $f \in \mathcal{K}$. For every $E \in \mathcal{S}$ we have $\varphi_E f_n \uparrow \varphi_E f$ a.e., therefore

$$\int \varphi_E f_n d\mu \uparrow \int \varphi_E f d\mu$$

by the monotone convergence theorem, that is, $(f_n \cdot \mu)(E) \uparrow (f \cdot \mu)(E)$; since $(f_n \cdot \mu)(E) \leq \nu(E)$ for all n , we conclude that $(f \cdot \mu)(E) \leq \nu(E)$, thus $f \in \mathcal{K}$.

(iv) According to (iii), there exists an $f \in \mathcal{K}$ such that

$$\int g d\mu \leq \int f d\mu \quad \text{for all } g \in \mathcal{K};$$

the proof will be completed by showing that $f \cdot \mu = \nu$. At any rate $f \cdot \mu \leq \nu$ (because $f \in \mathcal{K}$); writing $\rho = \nu - f \cdot \mu$, we have $\rho \geq 0$, thus ρ is a finite measure on \mathcal{S} . Since $\nu \ll \mu$ and $f \cdot \mu \ll \nu$, clearly $\rho \ll \mu$; our problem is to show that $\rho = 0$.

Assume to the contrary that $\rho \neq 0$. By the preceding lemma, there exists a μ -integrable function $h \geq 0$ such that $h \cdot \mu \leq \rho$ and $h \cdot \mu \neq 0$, that is, $0 \neq h \cdot \mu \leq \nu - f \cdot \mu$; then $f + h \geq 0$ and

$$(f + h) \cdot \mu = f \cdot \mu + h \cdot \mu \leq \nu,$$

so $f + h \in \mathcal{K}$. By the choice of f ,

$$\int (f + h) d\mu \leq \int f d\mu;$$

since $h \geq 0$, this implies that

$$\int f d\mu \leq \int f d\mu + \int h d\mu = \int (f + h) d\mu \leq \int f d\mu$$

whence equality throughout and $\int h d\mu = 0$. Thus $(h \cdot \mu)(X) = 0$, consequently $h \cdot \mu = 0$, a contradiction. \diamond

4.8.12. Corollary. Assume that μ is a finite measure on \mathcal{S} and let ν be a finite signed measure on \mathcal{S} . Then:

$$\nu \ll \mu \Leftrightarrow \nu = f \cdot \mu \text{ for some } f \in \mathcal{L}^1(\mu).$$

Proof. \Leftarrow : Example 4.8.7, (i).

\Rightarrow : By the Hahn decomposition (4.8.8) there exists a set $A \in \mathcal{S}$ such that the set functions $\rho = \nu_A$ and $\tau = -\nu_{X-A}$ are measures. Since $\rho \ll \mu$ and $\tau \ll \mu$ (4.8.7, (ii)), by the preceding theorem we have $\rho = g \cdot \mu$ and $\tau = h \cdot \mu$ for suitable μ -integrable functions g and h , therefore

$$\nu = \rho - \tau = g \cdot \mu - h \cdot \mu = (g - h) \cdot \mu$$

and $f = g - h$ meets the requirements. \diamond

Exercises

1. Let ν be a finite signed measure on \mathcal{S} and let (A_n) be a sequence of measurable sets. If $A_n \uparrow A$ or $A_n \downarrow A$ then $\nu_{A_n} \rightarrow \nu_A$ pointwise on \mathcal{S} .

2. If ν is a finite signed measure on \mathcal{S} , then ν is a bounded function on \mathcal{S} . {Hint: Hahn decomposition.}

3. The measures in the Hahn decomposition of a finite signed measure ν are unique: If A and B are measurable sets such that

$$\nu_A \geq 0, \nu_{X-A} \leq 0 \quad \text{and} \quad \nu_B \geq 0, \nu_{X-B} \leq 0,$$

then $\nu_A = \nu_B$ and $\nu_{X-A} = \nu_{X-B}$.

{Hint: $\nu_{A \cap B'} = (\nu_A)_{B'} = (\nu_{B'})_A$ is both ≥ 0 and ≤ 0 .}

4. With notations as in Exercise 3, one defines $|\nu| = \nu_A - \nu_{X-A}$, called the *total variation* of ν (it is a measure on \mathcal{S}). Then $\nu \ll \mu \Leftrightarrow |\nu| \ll \mu$.

5. Since the absolute convergence in 4.8.3, (i) is not used for the proof of the Hahn decomposition theorem, it can also be deduced from that theorem.

CHAPTER 5

Differentiation

- §5.1. Bounded variation, absolute continuity
- §5.2. Lebesgue's representation of AC functions
- §5.3. \limsup , \liminf of functions; Dini derivatives
- §5.4. Criteria for monotonicity
- §5.5. Semicontinuity
- §5.6. Semicontinuous approximations of Integrable Functions
- §5.7. F. Riesz's "Rising sun lemma"
- §5.8. Growth estimates of a continuous increasing function
- §5.9. Indefinite integrals are a.e. primitives
- §5.10. Lebesgue's "Fundamental theorem of calculus"
- §5.11. Measurability of derivatives of a monotone function
- §5.12. Lebesgue decomposition of a BV function
- §5.13. Lebesgue's criterion for Riemann-integrability

The focus of the chapter is on the "Fundamental theorem of calculus" for the Lebesgue theory, analogous to, but much harder than, the classical theorem of that name for the Riemann integral of a continuous function (the precise statements will be given shortly).

Following E. J. McShane¹, the basic strategy is to exploit the regularity of Lebesgue measure. The essential idea: (a) the "Fundamental theorem" is easy for the case of a continuous function; (b) every measurable set is approximable by open sets (whose characteristic functions are 'lower semicontinuous'); (c) every integrable function is approximable by simple functions, hence is approximable by lower semicontinuous functions; (d) the wisp of continuity in (c) facilitates the proof of the "Fundamental theorem" for functions that are not necessarily continuous. All in all, a lovely application of "Littlewood's second principle" (every measurable function is nearly continuous).²

¹E. J. McShane, *Integration* [Princeton University Press, Princeton, N.J., 1944], pp. 188-208.

²H. L. Royden, *Real analysis* [3rd edn., Macmillan, New York, 1988], Chapter 3, §6.

To simplify the following statements, let us say that a function $F : [a, b] \rightarrow \mathbb{R}$ is a *primitive* for a function $f : [a, b] \rightarrow \mathbb{R}$ if F is differentiable on (a, b) and $F'(x) = f(x)$ for all $x \in (a, b)$. {Such a function F is of course continuous on (a, b) , but nothing is said about its behavior at the endpoints.}

In the context of the **Riemann integral**, the classical “Fundamental theorem of calculus” can be expressed as follows:

If $f : [a, b] \rightarrow \mathbb{R}$ is continuous, then

(1) the function $F : [a, b] \rightarrow \mathbb{R}$ defined by the formula

$$F(x) = \int_a^x f(t)dt \quad (x \in [a, b])$$

is a **continuous primitive** for f ; moreover,

(2) every continuous primitive $G : [a, b] \rightarrow \mathbb{R}$ for f differs from F by a constant, hence can be used to calculate the Riemann integral of f :

$$\int_a^b f(t)dt = G(b) - G(a).$$

{The essential new ingredient needed for statement (2) is the Mean Value Theorem.}

The set of primitives that arise in (1) are precisely the functions F that are continuously differentiable on $[a, b]$ (one-sided at the endpoints) and vanish at a .

If the continuous function $f : [a, b] \rightarrow \mathbb{R}$ is replaced by an arbitrary Riemann-integrable function, what survives of (1) and (2)? {An answer, not entirely satisfactory, is given in §5.13, using results from the Lebesgue theory.}

The corresponding situation for the Lebesgue integral is, as we shall see in this chapter, satisfactory on all counts. Call a function $F : [a, b] \rightarrow \mathbb{R}$ an *a.e.-primitive* for a function $f : [a, b] \rightarrow \mathbb{R}$ if $F' = f$ a.e., in the sense that, for almost every $x \in (a, b)$, F is differentiable at x and $F'(x) = f(x)$. The “Fundamental theorem of calculus” for the **Lebesgue integral** takes the following form:

If $f : [a, b] \rightarrow \mathbb{R}$ is Lebesgue-integrable, then

(1) the function $F : [a, b] \rightarrow \mathbb{R}$ defined by the formula

$$F(x) = \int_{[a, x]} f d\lambda \quad (x \in [a, b])$$

is an **absolutely continuous**³ a.e.-primitive for f ; moreover,

³ Cf. 5.1.9 below.

(2) every absolutely continuous a.e.-primitive $G : [a, b] \rightarrow \mathbb{R}$ for f differs from F by a constant, hence can be used to calculate the Lebesgue integral of f :

$$\int_a^b f \, d\lambda = G(b) - G(a).$$

The classes of functions that appear in the above theorem can be characterized succinctly without reference to integration theory:

(i) The functions F that arise in (1) are precisely the absolutely continuous functions on $[a, b]$ such that $F(a) = 0$.

(ii) A function $f : [a, b] \rightarrow \mathbb{R}$ is Lebesgue-integrable if and only if there exists an absolutely continuous function $G : [a, b] \rightarrow \mathbb{R}$ such that $G' = f$ a.e. (It is implicit here that every absolutely continuous function has a derivative at almost every point of (a, b) .⁴)

In the classical “Fundamental theorem” (with f continuous), the proofs of (1) and (2) are easy⁵; for the Lebesgue case, they are challenging.⁶

What can be said for the “intermediate” case of a Riemann-integrable function? As we shall see in §5.13, every Riemann-integrable function $f : [a, b] \rightarrow \mathbb{R}$ is Lebesgue-integrable and the two concepts of integral coincide for f ; consequently, the statements (1) and (2) of the “Fundamental theorem” of the Lebesgue theory apply in particular to a Riemann-integrable function f . The main result in §5.13 is Lebesgue’s characterization of Riemann-integrability: A bounded function f on $[a, b]$ is Riemann-integrable if and only if its set of discontinuities has (Lebesgue) measure zero. A useful characterization of indefinite integrals (of Riemann-integrable functions) appears to be elusive.⁷

5.1. Bounded Variation, Absolute Continuity

5.1.1. *Definition.* Let $f : [a, b] \rightarrow \mathbb{R}$, $a < b$. For every subdivision

$$\sigma = \{a = x_0 < x_1 < \cdots < x_n = b\}$$

of $[a, b]$ into subintervals $[x_{k-1}, x_k]$, let us write

$$|f(\sigma)| = \sum_{k=1}^n |f(x_k) - f(x_{k-1})|;$$

⁴ Cf. 5.9.4 below.

⁵ *First course*, p. 151, 9.4.6.

⁶ Cf. §5.10 below.

⁷ Cf. the author, “Why there is no ‘Fundamental theorem of calculus’ for the Riemann integral”, *Expositiones Mathematicae* 11 (1993), 271–279.

the supremum of all such sums is called the **total variation** of f on $[a, b]$, denoted $V_a^b f$, thus

$$V_a^b f = \sup \{ |f(\sigma)| : \sigma \text{ a subdivision of } [a, b] \}.$$

Obviously $V_a^b f \geq 0$ (possibly $= +\infty$).

If $a \leq c < d \leq b$, we write $V_c^d f$ for the total variation of the restriction of f to $[c, d]$, that is,

$$V_c^d f = V_c^d(f|_{[c, d]}).$$

Convention: $V_a^a f = 0$. If $V_a^b f < +\infty$ then f is said to be of **bounded variation** (briefly, BV) on $[a, b]$.

One of the main goals of the chapter is to prove that a function of bounded variation is differentiable almost everywhere (5.12.8).

5.1.2. *Remarks.* (i) If $f : [a, b] \rightarrow \mathbb{R}$ is increasing then $|f(\sigma)| = f(b) - f(a)$ for every subdivision σ of $[a, b]$ (in the formula of Definition 5.1.1, the absolute value signs on the right side can be omitted and the sum telescopes), therefore f is BV and $V_a^b f = f(b) - f(a)$.

(ii) If f satisfies a Lipschitz condition $|f(x) - f(y)| \leq K|x - y|$ on $[a, b]$, then f is BV and $V_a^b f \leq K(b - a)$.

(iii) If $f : [a, b] \rightarrow \mathbb{R}$, $[c, d]$ is a subinterval of $[a, b]$ and $g = f|_{[c, d]}$, then $V_c^d g \leq V_a^b f$ (because every subdivision of $[c, d]$ is part of a subdivision of $[a, b]$); in other words, $V_c^d f \leq V_a^b f$. In particular, f BV $\Rightarrow g$ BV (and we say that f is BV on $[c, d]$).

(iv) $|f(b) - f(a)| \leq V_a^b f$ for every function $f : [a, b] \rightarrow \mathbb{R}$ (consider the trivial subdivision $\sigma = \{a = x_0 < x_1 = b\}$). For every $x \in [a, b]$,

$$|f(x)| \leq |f(x) - f(a)| + |f(a)| \leq V_a^x f + |f(a)| \leq V_a^b f + |f(a)|;$$

writing $\|f\|_\infty = \sup \{ |f(x)| : x \in [a, b] \}$ (possibly $+\infty$), we have

$$\|f\|_\infty \leq V_a^b f + |f(a)|.$$

The inequality is of interest only when f is of bounded variation, and shows that f BV $\Rightarrow f$ bounded.

(v) $V_a^b f = 0 \Leftrightarrow f$ is constant.

{Proof: If $V_a^b f = 0$ then $|f(x) - f(a)| \leq V_a^x f \leq V_a^b f = 0$ for all $x \in [a, b]$, therefore f is constant. The reverse implication is clear from Definition 5.1.1.}

The correspondence $\sigma \mapsto |f(\sigma)|$ is monotone in an appropriate sense:

5.1.3. Lemma. Let $f : [a, b] \rightarrow \mathbb{R}$ and let σ, τ be subdivisions of $[a, b]$. Then

$$\sigma \prec \tau \Rightarrow |f(\sigma)| \leq |f(\tau)|.$$

Proof. Here $\sigma \prec \tau$ means that τ is a refinement of σ (also written $\tau \succ \sigma$), that is, every point of σ is also a point of τ . The assertion is that

$\sigma \mapsto |f(\sigma)|$ is an increasing function on the set of subdivisions (ordered by \prec).

It suffices to consider the case that τ is obtained from σ by adding one new point c . Say $\sigma = \{a = x_0 < x_1 < \cdots < x_n = b\}$ and $x_{j-1} < c < x_j$. Then

$$(*) \quad |f(x_j) - f(x_{j-1})| \leq |f(x_j) - f(c)| + |f(c) - f(x_{j-1})|;$$

the left side of $(*)$ is the j 'th term of $|f(\sigma)|$, the right side is the sum of the j 'th and $(j+1)$ 'th terms of $|f(\tau)|$, and the remaining terms of $|f(\sigma)|$ are identical with the remaining terms of $|f(\tau)|$, thus $|f(\sigma)| \leq |f(\tau)|$ is immediate from $(*)$. \diamond

In analogy with monotone sequences, it is suggestive to write $|f(\sigma)| \uparrow V_a^b f$ as $\sigma \uparrow \infty$, where " $\sigma \uparrow \infty$ " symbolizes indefinite refinement; more precisely, for every $r < V_a^b f$, there exists a subdivision σ_0 such that $r < |f(\sigma_0)| \leq V_a^b f$ (whence the same inequalities hold with σ_0 replaced by any subdivision σ such that $\sigma \succ \sigma_0$).

A key property of total variation is its 'additivity' in the following sense:

5.1.4. Theorem. If $f : [a, b] \rightarrow \mathbb{R}$ and $a \leq c \leq b$, then $V_a^b f = V_a^c f + V_c^b f$.

Proof. Note that infinite values are allowed. The asserted equality holds trivially if c coincides with one of the endpoints, or if one of $V_a^c f$, $V_c^b f$ is infinite (5.1.2, (iii)), thus we can suppose that $a < c < b$ and that f is BV on both $[a, c]$ and $[c, b]$. The equality will be proved by showing that each side is \leq the other.

Proof of \leq : If σ is any subdivision of $[a, b]$, it will suffice to show that

$$|f(\sigma)| \leq V_a^c f + V_c^b f.$$

Let τ be a refinement of σ that includes c as one of its points. The points of τ that are in $[a, c]$ determine a subdivision τ_1 of $[a, c]$, and those that are in $[c, b]$ determine a subdivision τ_2 of $[c, b]$. Writing $|f(\tau_1)|$ for $|g(\tau_1)|$, where g is the restriction of f to $[a, c]$, and similarly for $|f(\tau_2)|$, it is clear that $|f(\tau)| = |f(\tau_1)| + |f(\tau_2)|$; citing the lemma, we have

$$|f(\sigma)| \leq |f(\tau)| = |f(\tau_1)| + |f(\tau_2)| \leq V_a^c f + V_c^b f,$$

as we wished to show. The inequality $V_a^b f \leq V_a^c f + V_c^b f$ just established shows in particular that f is also BV on $[a, b]$.

Proof of \geq : Let σ_1 be any subdivision of $[a, c]$, σ_2 any subdivision of $[c, b]$. If σ is the subdivision of $[a, b]$ obtained by joining σ_1 and σ_2 at their common endpoint, then $|f(\sigma)| = |f(\sigma_1)| + |f(\sigma_2)|$, thus

$$(*) \quad |f(\sigma_1)| = |f(\sigma)| - |f(\sigma_2)| \leq V_a^b f - |f(\sigma_2)|;$$

this shows that for each σ_2 , $V_a^b f - |f(\sigma_2)|$ is an upper bound for all the $|f(\sigma_1)|$, therefore

$$V_a^c f \leq V_a^b f - |f(\sigma_2)|,$$

that is,

$$|f(\sigma_2)| \leq V_a^b f - V_a^c f$$

(all numbers in sight are finite, so transposition poses no problem); since the last inequality holds for every σ_2 , we have

$$V_c^b f \leq V_a^b f - V_a^c f$$

and another transposition finishes the job. \diamond

5.1.5. Corollary. If $f : [a, b] \rightarrow \mathbb{R}$ and $a < c < b$, then

$$f \text{ is BV on } [a, b] \Leftrightarrow f \text{ is BV on } [a, c] \text{ and } [c, b].$$

Proof. The left side of $V_a^b f = V_a^c f + V_c^b f$ is finite if and only if both terms on the right side are finite. \diamond

There is a useful 'algebra' of the bounded variation property:

5.1.6. Theorem. If $f : [a, b] \rightarrow \mathbb{R}$ and $g : [a, b] \rightarrow \mathbb{R}$ are BV, then so are $f + g$, cf ($c \in \mathbb{R}$), $|f|$, $f \cup g$, $f \cap g$ and fg .

Proof. By (iv) of 5.1.2, f and g are bounded. For all x, y in $[a, b]$,

$$(fg)(x) - (fg)(y) = f(x)[g(x) - g(y)] + [f(x) - f(y)]g(y),$$

therefore

$$|(fg)(x) - (fg)(y)| \leq \|f\|_\infty |g(x) - g(y)| + \|g\|_\infty |f(x) - f(y)|;$$

it follows that, for every subdivision σ of $[a, b]$,

$$|(fg)(\sigma)| \leq \|f\|_\infty |g(\sigma)| + \|g\|_\infty |f(\sigma)|,$$

thus fg is BV and

$$V_a^b(fg) \leq \|f\|_\infty V_a^b g + \|g\|_\infty V_a^b f.$$

Similarly, the easily verified relations

$$|(f+g)(\sigma)| \leq |f(\sigma)| + |g(\sigma)|, \quad |(cf)(\sigma)| = |c| \cdot |f(\sigma)|, \quad ||f|(\sigma)| \leq |f(\sigma)|$$

show that $f + g$, cf and $|f|$ are BV, with

$$V_a^b(f+g) \leq V_a^b f + V_a^b g, \quad V_a^b(cf) = |c| \cdot V_a^b f, \quad V_a^b|f| \leq V_a^b f.$$

The assertions about $f \cup g$ and $f \cap g$ follow from the formulas

$$f \cup g = \frac{1}{2}(f + g + |f - g|), \quad f \cap g = \frac{1}{2}(f + g - |f - g|)$$

and what has already been shown. \diamond

A nice application of the concept of total variation is a ‘structure theorem’ for functions of bounded variation:

5.1.7. Theorem. *The following conditions on a function $f : [a, b] \rightarrow \mathbb{R}$ are equivalent:*

(a) f is BV;

(b) f is the difference of two increasing functions.

Proof. (b) \Rightarrow (a): If $f = g - h$ with g and h increasing, then f is BV because it is a linear combination of BV’s (5.1.2, (i) and 5.1.6). {Incidentally, g and h are not uniquely determined by f ; for example, if f is increasing, then $f = f - 0 = 2f - f$.}

(a) \Rightarrow (b): The proof will exhibit a ‘canonical’ (same procedure for all f) representation $f = p - n$ of f as a difference of increasing functions. For each $x \in [a, b]$, $V_a^x f$ is a real number (5.1.2, (iii)); define $p : [a, b] \rightarrow \mathbb{R}$ by the formula

$$p(x) = V_a^x f \quad (x \in [a, b])$$

(a definition reminiscent of ‘indefinite integrals’). Thus $p(a) = 0$ and $p(b) = V_a^b f$. Defining $n = p - f$, we have $f = p - n$ and it remains only to show that p and n are increasing. If $a \leq x < y \leq b$ then, by 5.1.4,

$$(*) \quad p(y) = V_a^y f = V_a^x f + V_x^y f = p(x) + V_x^y f \geq p(x),$$

thus p is increasing; moreover, by (*),

$$\begin{aligned} n(y) - n(x) &= [p(y) - f(y)] - [p(x) - f(x)] \\ &= [p(y) - p(x)] - [f(y) - f(x)] \\ &= V_x^y f - [f(y) - f(x)], \end{aligned}$$

where $f(y) - f(x) \leq |f(y) - f(x)| \leq V_x^y f$ (5.1.2, (iv)), thus $n(y) - n(x) \geq 0$. \diamond

5.1.8. *Definition.* With notations as in the preceding proof, the formula $f = p - n$ is called the **Jordan decomposition** of the function f of bounded variation.

The “Fundamental theorem of Calculus” of the Lebesgue theory, to be proved later in the chapter, requires a condition stronger than BV (called absolute continuity). Before defining this concept, some notational preparation is in order.

If $f : [a, b] \rightarrow \mathbb{R}$ and $\sigma = \{a = x_0 < x_1 = b\}$ is the trivial subdivision of $[a, b]$ then, according to 5.1.1, $|f(\sigma)| = |f(b) - f(a)|$. If I is a subinterval of $[a, b]$ with endpoints c, d (whose ordering is not specified), it will be useful to have a notation for $|f(c) - f(d)|$; I propose writing

$$|f(I)| = |f(c) - f(d)|.$$

The notation is not entirely satisfactory—we must resist the temptation

to interpret $f\langle I \rangle$ as the image $f(I)$ of the set I under the function f —but I could not think of a better one (cf. 5.1.18). At least, the notation is compatible with 5.1.1: if $\sigma = \{a = x_0 < x_1 < \cdots < x_n = b\}$ and $I_k = [x_{k-1}, x_k]$, then

$$|f(\sigma)| = \sum_{k=1}^n |f\langle I_k \rangle|.$$

5.1.9. Definition. A function $f : [a, b] \rightarrow \mathbb{R}$ is said to be **absolutely continuous** (briefly, AC) if it has the following property: for every $\epsilon > 0$ there exists a $\delta > 0$ such that, for **nonoverlapping**⁸ subintervals I_1, \dots, I_n of $[a, b]$, where I_k has endpoints a_k, b_k with $a_k \leq b_k$ ($k = 1, \dots, n$), we have

$$\sum_{k=1}^n (b_k - a_k) \leq \delta \Rightarrow \sum_{k=1}^n |f(b_k) - f(a_k)| \leq \epsilon,$$

concisely, $\sum_{k=1}^n \lambda(I_k) \leq \delta \Rightarrow \sum_{k=1}^n |f\langle I_k \rangle| \leq \epsilon$.

5.1.10. Remarks. (i) $AC \Rightarrow$ continuous (contemplate $n = 1$ in 5.1.9).

(ii) If f is AC on $[a, b]$, then the restriction of f to a subinterval $[c, d]$ is also AC (the δ that works for f will also work for its restriction). This is also expressed by saying that f is AC on $[c, d]$.

(iii) With notations as in 5.1.9, if $[c, d] \subset [a, b]$ and $d - c \leq \delta$ then $V_c^d f \leq \epsilon$.

{Proof: Every subdivision τ of $[c, d]$ determines a list of nonoverlapping intervals with total length $d - c \leq \delta$, therefore $|f(\tau)| \leq \epsilon$ by the choice of δ .}

(iv) $AC \Rightarrow BV$.

{Proof: Assuming $f : [a, b] \rightarrow \mathbb{R}$ absolutely continuous, we are to show that f is of bounded variation. Let $\epsilon = 1$ (for example) and choose $\delta > 0$ as in 5.1.9. Choose a subdivision $\sigma = \{a = x_0 < x_1 < \cdots < x_n = b\}$ with $x_k - x_{k-1} \leq \delta$ for all k . By (iii), f is BV on each $[x_{k-1}, x_k]$, therefore f is BV on $[a, b]$ by 5.1.5.}

(v) There exist BV functions—even continuous ones—that are not AC; Lebesgue's famous example is sketched at the end of the section.

(vi) Lipschitz $\Rightarrow AC$.

{Proof: If $K > 0$ is a constant such that $|f(x) - f(y)| \leq K|x - y|$ for all x, y in $[a, b]$, then $\delta = \epsilon/K$ 'works' in 5.1.9.}

The 'algebra' of absolute continuity can be derived using the same basic techniques as for bounded variation (5.1.6):

⁸Intervals I and J of \mathbb{R} are said to be nonoverlapping if $I \cap J$ is either \emptyset or a singleton, in other words, $\lambda(I \cap J) = 0$; an equivalent condition is that $I \cap J$ have empty interior. Intervals I_1, \dots, I_n are said to be nonoverlapping if they are 'pairwise nonoverlapping', that is, if I_j and I_k are nonoverlapping for all $j \neq k$.

5.1.11. Theorem. If $f : [a, b] \rightarrow \mathbb{R}$ and $g : [a, b] \rightarrow \mathbb{R}$ are AC, then so are $f + g$, cf ($c \in \mathbb{R}$), $|f|$, $f \cup g$, $f \cap g$ and fg .

5.1.12. Corollary. Every real polynomial function on $[a, b]$ is AC.

Proof. The identity function $x \mapsto x$ is AC, therefore so is every linear combination of its powers. \diamond

Here is an example with more meat on the bones:

5.1.13. Theorem. If $f : [a, b] \rightarrow \mathbb{R}$ is Lebesgue-integrable and if

$$F(x) = \int_a^x f \, d\lambda \quad (x \in [a, b])$$

then F is AC on $[a, b]$.

Proof. {It will be shown in the next section that every absolutely continuous function is of the form $F + \text{constant}$, for a suitable integrable function f .}

Recall that the symbol on the right side is just a suggestive notation for

$$\int_{[a, x]} f \, d\lambda = \int \varphi_{[a, x]} f \, d\lambda,$$

where $\varphi_{[a, x]}$ is the characteristic function of $[a, x]$. {It is not difficult to show that this is also equal to the Lebesgue integral of the restriction of f to $[a, x]$, with respect to Lebesgue measure on $[a, x]$ (start with the case that f is the characteristic function of a Lebesgue-measurable subset of $[a, b]$).}

Consider first the case that f is bounded, say $|f| \leq K$. Then F is Lipschitz: if $a \leq x \leq y \leq b$ then

$$|F(x) - F(y)| = \left| \int_{[x, y]} f \, d\lambda \right| \leq K|y - x|.$$

Thus F is AC by 5.1.10, (vi).

Now consider the general case. Writing $f = f^+ - f^-$, we can suppose that $f \geq 0$. Given any $\epsilon > 0$, choose a simple function g such that $0 \leq g \leq f$ and

$$\int f - \int g \leq \epsilon,$$

and define

$$G(x) = \int_a^x g \, d\lambda \quad (x \in [a, b]);$$

since f and g are nonnegative, F and G are increasing. By the first case considered, G is AC; choose $\delta > 0$ to go along with ϵ as in 5.1.9. Assuming I_1, \dots, I_n are nonoverlapping subintervals of $[a, b]$ with $\sum \lambda(I_k) \leq \delta$, it will suffice to show that $\sum |F(I_k)| \leq 2\epsilon$. If I_k has

endpoints $a_k \leq b_k$, then

$$|F\langle I_k \rangle| = F(b_k) - F(a_k) = \int_{a_k}^{b_k} f = \int_{I_k} f$$

and similarly

$$|G\langle I_k \rangle| = \int_{I_k} g.$$

Let $A = I_1 \cup \dots \cup I_n$; since singletons are negligible and indefinite integrals are additive set functions, it follows that

$$\begin{aligned} \sum_{k=1}^n |F\langle I_k \rangle| &= \sum_{k=1}^n \int_{I_k} f = \int_A f \\ &= \int_A (f - g) + \int_A g \\ &\leq \int_A (f - g) + \int_A g \leq \epsilon + \int_A g \\ &= \epsilon + \sum_{k=1}^n \int_{I_k} g = \epsilon + \sum_{k=1}^n |G\langle I_k \rangle| \leq \epsilon + \epsilon \end{aligned}$$

(the last inequality by the choice of δ). \diamond

The 'monotone constituents' of an absolutely continuous function are themselves absolutely continuous:

5.1.14. Theorem. If $f : [a, b] \rightarrow \mathbb{R}$ is AC (hence BV) and if $f = p - n$ is its Jordan decomposition (5.1.8), then p and n are also AC.

Proof. It is enough to show that p is AC (then $n = p - f$ will be AC by 5.1.11). Given any $\epsilon > 0$, choose $\delta > 0$ as in 5.1.9; assuming $I_k = [a_k, b_k]$ ($k = 1, \dots, r$) are nonoverlapping subintervals of $[a, b]$ such that $\sum (b_k - a_k) \leq \delta$, it suffices to show that $\sum |p\langle I_k \rangle| \leq 2\epsilon$.

For each index k , f is BV on I_k ; choose a subdivision σ_k of I_k such that

$$(*) \quad |f(\sigma_k)| > V_{a_k}^{b_k} f - \epsilon/r$$

(possible by the definition of total variation as a least upper bound). Then

$$\begin{aligned} \sum_{k=1}^r |p\langle I_k \rangle| &= \sum_{k=1}^r [p(b_k) - p(a_k)] \\ &= \sum_{k=1}^r [V_a^{b_k} f - V_a^{a_k} f] = \sum_{k=1}^r V_{a_k}^{b_k} f \\ &< \sum_{k=1}^r [|f(\sigma_k)| + \epsilon/r] = \epsilon + \sum_{k=1}^r |f(\sigma_k)| \end{aligned}$$

(the steps are justified by the monotonicity and defining formula of p , Theorem 5.1.4, and the inequalities (*)); thus $\sum |p\langle I_k \rangle| < \epsilon + \sum |f(\sigma_k)| \leq$

$\epsilon + \epsilon$ by the choice of δ (the sum of the lengths of the subintervals making up the various σ_k is $\sum(b_k - a_k) \leq \delta$). \diamond

On the way to proving that an absolutely continuous function maps negligible sets to negligible sets:

5.1.15. Lemma. *If $f : [a, b] \rightarrow \mathbb{R}$ is absolutely continuous and $f = p - n$ is its Jordan decomposition, then, for every closed subinterval I of $[a, b]$,*

$$\lambda(f(I)) \leq \lambda(p(I)) + \lambda(n(I)).$$

Proof. By the Intermediate Value Theorem, $f(I) = [f(r), f(s)]$ for suitable r, s in I . Let J be the closed subinterval of I with endpoints r, s (we need not know which is the larger). Then

$$\begin{aligned} \lambda(f(I)) &= f(s) - f(r) = [p(s) - n(s)] - [p(r) - n(r)] \\ &= [p(s) - p(r)] - [n(s) - n(r)] \\ &\leq |p(s) - p(r)| + |n(s) - n(r)|. \end{aligned}$$

Since p is monotone and continuous (5.1.14), $p(J)$ is the closed interval with endpoints $p(r), p(s)$, therefore $\lambda(p(J)) = |p(s) - p(r)|$; similarly $\lambda(n(J)) = |n(s) - n(r)|$, therefore (by the earlier inequality and the inclusion $J \subset I$)

$$\lambda(f(I)) \leq \lambda(p(J)) + \lambda(n(J)) \leq \lambda(p(I)) + \lambda(n(I)). \quad \diamond$$

5.1.16. Theorem. *If $f : [a, b] \rightarrow \mathbb{R}$ is AC and N is a negligible subset of $[a, b]$, then $f(N)$ is also negligible.*

Proof. Suppose first that f is increasing. For every subinterval $I = [c, d]$ of $[a, b]$, we have $f(I) = [f(c), f(d)]$ (by monotonicity and the Intermediate Value Theorem). Given any $\epsilon > 0$, choose $\delta > 0$ as in 5.1.9. Since N is negligible, there exists a sequence of intervals $I_k = [a_k, b_k]$ such that $N \subset \bigcup I_k$ and $\sum(b_k - a_k) \leq \delta$. Replacing I_k by $I_k \cap [a, b]$, we can suppose that $I_k \subset [a, b]$. Then

$$f(N) \subset \bigcup f(I_k) = \bigcup [f(a_k), f(b_k)],$$

therefore

$$(*) \quad \lambda^*(f(N)) \leq \sum [f(b_k) - f(a_k)].$$

We can suppose further that the I_k are nonoverlapping. {Proof: First ‘disjointify’ by defining $A_k = [a_1, b_1] \cup \dots \cup [a_k, b_k]$, then $B_1 = A_1$, $B_{k+1} = A_{k+1} - A_k$, so that $\bigcup [a_k, b_k]$ is expressed as a disjoint union $\bigcup B_k$; write each B_k as a finite disjoint union of intervals (of the four possible kinds; cf. 4.6.2), then restore all missing endpoints.} For each positive integer r , the intervals I_1, \dots, I_r are nonoverlapping and $\sum_1^r (b_k - a_k) \leq \delta$, therefore $\sum_1^r [f(b_k) - f(a_k)] \leq \epsilon$ by the choice of δ ; since r is arbitrary,

$\sum [f(b_k) - f(a_k)] \leq \epsilon$, thus $\lambda^*(f(N)) \leq \epsilon$ by (*). This proves the corollary for an increasing AC function.

In the general case, let $f = p - n$ be the Jordan decomposition of f ; by 5.1.14, both p and n are AC. Given any $\epsilon > 0$, choose $\delta > 0$ in 5.1.9 to 'work' for both p and n . With the notations $N \subset \bigcup I_k$ as in the first part of the proof, we have $f(N) \subset \bigcup f(I_k)$, therefore $\lambda^*(f(N)) \leq \sum \lambda(f(I_k))$. By the lemma, $\lambda(f(I_k)) \leq \lambda(p(I_k)) + \lambda(n(I_k))$, thus

$$(**) \quad \lambda^*(f(N)) \leq \sum \lambda(p(I_k)) + \sum \lambda(n(I_k)).$$

If $I_k = [a_k, b_k]$ then $p(I_k) = [p(a_k), p(b_k)]$ and $n(I_k) = [n(a_k), n(b_k)]$, thus the inequality (**) may be written

$$\lambda^*(f(N)) \leq \sum [p(b_k) - p(a_k)] + \sum [n(b_k) - n(a_k)];$$

the first part of the proof also shows that each sum on the right is $\leq \epsilon$, whence $\lambda^*(f(N)) \leq 2\epsilon$. \diamond

5.1.17. *Lebesgue's singular function*⁹. I recommend skipping the rest of the section if you have access to McShane's *Integration* (pp. 48-50); you will not find a more lucid explanation of Lebesgue's function anywhere (from beginning to end, the discussion occupies less than a page and a half and explains *everything*).

The objective is to construct an increasing (hence BV) continuous function $f : [0, 1] \rightarrow [0, 1]$ that is not AC. The function is paradoxical in that its graph is 'almost always horizontal', yet manages to climb continuously (no jumps) from 0 to 1.

The construction begins by defining a function $f_0 : [0, 1] - \Gamma \rightarrow [0, 1]$ on the complement of the Cantor set Γ (§2.3); one then defines $f : [0, 1] \rightarrow [0, 1]$ by extrapolating from the values of f_0 on $[0, 1] - \Gamma$. Figure 1 shows the first three steps of the construction of f_0 .

Step 1: Define f_0 to be $\frac{1}{2}$ on the open middle third $(\frac{1}{3}, \frac{2}{3})$ of the unit interval (the first subinterval that was expelled in the construction of the Cantor set).

Step 2: Define f_0 to be $\frac{1}{4}$ on $(\frac{1}{9}, \frac{2}{9})$, and $\frac{3}{4}$ on $(\frac{7}{9}, \frac{8}{9})$ (the subintervals expelled in the second step of the construction of the Cantor set).

Step 3: Define f_0 to be $\frac{1}{8}$ on $(\frac{1}{27}, \frac{2}{27})$; $\frac{3}{8}$ on $(\frac{7}{27}, \frac{8}{27})$; $\frac{5}{8}$ on $(\frac{19}{27}, \frac{20}{27})$; and $\frac{7}{8}$ on $(\frac{25}{27}, \frac{26}{27})$.

'And so on by induction' (but it is a little messy to get it all down on paper!)¹⁰; after n steps the function f_0 has been defined on $2^n - 1$ open intervals.

We now have a function $f_0 : [0, 1] - \Gamma \rightarrow [0, 1]$ defined on a disjoint union of (countably many) open intervals. The function $f : [0, 1] \rightarrow [0, 1]$,

⁹ The meaning of "singular" is explained in §5.12 (specifically, Definition 5.12.10).

¹⁰ The notational problems are vanquished in E. Hewitt and K. Stromberg's *Real and abstract analysis* [Springer, New York, 1965], p. 113, Exercise 8.2.8.

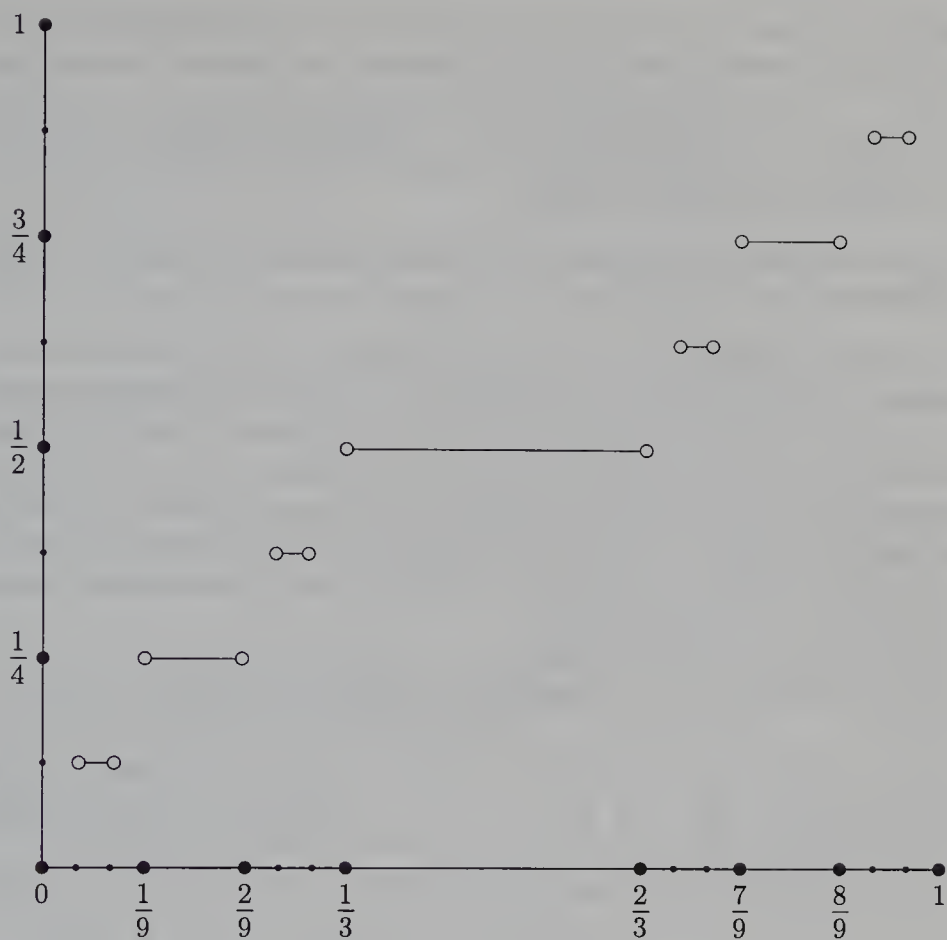


Figure 1

defined by $f(0) = 0$ and by the formula

$$f(x) = \sup\{f_0(t) : t \in [0, 1] - \Gamma, t \leq x\}$$

for $x \in (0, 1]$, extends f_0 and has all the required properties [McShane, loc.cit.].

5.1.18. *Rationale for a notation.* For a bounded interval I , think of the notation $|f(I)|$ in two steps: write $\langle I \rangle = \{c, d\}$ for the set of endpoints of I , let $f(\langle I \rangle) = \{f(c), f(d)\}$ be the image of $\{c, d\}$ under f , and, for every doublet $\{x, y\} \subset \mathbb{R}$, write $|\{x, y\}| = |x - y|$; then

$$|f(c) - f(d)| = |\{f(c), f(d)\}| = |f(\langle I \rangle)|$$

and we are proposing to abbreviate $|f(\langle I \rangle)|$ to $|f(I)|$. Incidentally, $|\langle I \rangle| = \lambda(I)$, the length (or Lebesgue measure) of I .

Exercises

1. With notations as in the proof of 5.1.6, the relations $V(f + g) \leq Vf + Vg$, $V(cf) = |c| \cdot Vf$, $V|f| \leq Vf$ and $V(fg) \leq \|f\|_\infty \cdot Vg + \|g\|_\infty \cdot Vf$

hold for *all* functions $f, g : [a, b] \rightarrow \mathbb{R}$ (not necessarily BV). {For the second and fourth relations, recall the convention $0 \cdot (+\infty) = 0$ (1.15.4).}

2. If $f : [a, b] \rightarrow \mathbb{R}$ is BV and if f is bounded away from 0, then $1/f$ is also BV. {The assumption is that there is a constant $k > 0$ such that $|f(x)| \geq k$ for all $x \in [a, b]$.}

3. A function $f : [a, b] \rightarrow \mathbb{R}$ is BV if and only if f^+ and f^- are BV. {Recall that $f^+ = f \cup 0$ and $f^- = -(f \cap 0)$.}

4. (i) The set $\mathcal{BV} = \mathcal{BV}[a, b]$ of all functions $f : [a, b] \rightarrow \mathbb{R}$ of bounded variation is a subalgebra of the algebra $\mathcal{B} = \mathcal{B}[a, b]$ of all bounded functions on $[a, b]$, equipped with the pointwise operations $f + g$, cf , fg .

{An *algebra* is a ring with a vector space structure (sharing the same addition), products and scalar multiples being 'compatible' in the sense that $(cx)y = c(xy) = x(cy)$ for all ring elements x, y and all scalars c .}

(ii) For a function $f : [a, b] \rightarrow \mathbb{R}$, $V_a^b f = 0 \Leftrightarrow f$ is constant, therefore $f = 0 \Leftrightarrow f(a) = 0$ and $V_a^b f = 0$.

(iii) For every $f \in \mathcal{BV}$ define $\|f\| = |f(a)| + V_a^b f$. For all $f, g \in \mathcal{BV}$ and $c \in \mathbb{R}$, $\|f + g\| \leq \|f\| + \|g\|$, $\|cf\| = |c| \cdot \|f\|$, and $\|f\| = 0 \Leftrightarrow f = 0$.

(iv) For every $f \in \mathcal{BV}$, $\|f\|_\infty \leq \|f\|$; if (f_n) is a sequence in \mathcal{BV} such that $\|f_m - f_n\| \rightarrow 0$ as $m, n \rightarrow \infty$, then there exists an $f \in \mathcal{BV}$ such that $\|f_n - f\| \rightarrow 0$.¹¹

(v) For every $f \in \mathcal{BV}$, define $\|f\|' = \|f\|_\infty + V_a^b f$. Then $\|\cdot\|'$ has the properties listed for $\|\cdot\|$ in (iii) and (iv). Moreover, $\|fg\|' \leq \|f\|' \|g\|'$ for all $f, g \in \mathcal{BV}$.

{Hint: To prove the analogue of (iv) for $\|\cdot\|'$, note that

$$\|f\|_\infty \leq \|f\| \leq \|f\|' \quad \text{and} \quad V_a^b f \leq \|f\|,$$

then infer the assertion for $\|\cdot\|'$ from the assertion for $\|\cdot\|$. For the inequality concerning fg , inspect the proof of 5.1.6.}

5. (i) If $f : [a, b] \rightarrow \mathbb{R}$ is BV and continuous, and if $f = p - n$ is its Jordan decomposition, then p and n are also continuous.¹²

(ii) The conclusion of 5.1.15 is true for f BV and continuous.

6. Let $f : [a, b] \rightarrow \mathbb{R}$, $a < c < b$. Then:

(i) f is AC on $[a, b]$ if and only if it is AC on $[a, c]$ and $[c, b]$.

(ii) f is AC if and only if f^+ and f^- are AC.

(iii) If f is AC and bounded away from 0 (cf. Exercise 2) then $1/f$ is also AC.

¹¹ T.H. Hildebrandt, *Introduction to the theory of integration* [Academic Press, New York, 1963], p. 43, Theorem 8.6.

¹² Cf. B. Sz.-Nagy, *Introduction to real functions and orthogonal expansions* [Oxford, New York, 1965], p. 97, Theorem, or T.H. Hildebrandt, *op. cit.*, p. 40.

7. (i) The set $\mathcal{AC} = \mathcal{AC}[a, b]$ of all AC functions $f : [a, b] \rightarrow \mathbb{R}$ is a subalgebra of the algebra $\mathcal{C} = \mathcal{C}[a, b]$ of all continuous functions on $[a, b]$ (equipped, as in Exercise 4, with the pointwise operations).

(ii) There exists a sequence of AC functions (f_n) such that $\|f_m - f_n\|_\infty \rightarrow 0$ as $m, n \rightarrow \infty$, for which no AC function f exists such that $\|f_n - f\|_\infty \rightarrow 0$.

{Hint: Borrowing from the future (§6.9), there exists a sequence of polynomial functions that converges uniformly to Lebesgue's singular function.}

8. With notations as in 5.1.9 (in particular, f absolutely continuous), $\sum \lambda(I_k) \leq \delta$ need not imply $\sum |f(I_k)| \leq \epsilon$ if the intervals I_k are permitted to overlap.

{Hint: Define $f : [0, 1] \rightarrow \mathbb{R}$ by $f(0) = 0$ and $f(t) = \frac{1}{2}t^{-1/2}$ for $t \in (0, 1]$, and let F be its (absolutely continuous) indefinite integral. Let $0 < \epsilon < 1$, choose $\delta > 0$ as in 5.1.9, let n be a positive integer such that $1/n < \delta$, and contemplate repeating the interval $[0, 1/n^2]$ n times.}

9. Let us say that a function $f : [a, b] \rightarrow \mathbb{R}$ is *negligent* (with respect to Lebesgue measure) if it has the property in the conclusion of 5.1.16: $N \subset [a, b]$ negligible $\Rightarrow f(N)$ negligible.

(i) (Banach) A function f is AC if and only if it is continuous, BV and negligent.¹³

(ii) f is AC if and only if it can be written as a difference $f = g - h$ with g and h continuous, strictly increasing and negligent.¹⁴

5.2. Lebesgue's Representation of Absolutely Continuous Functions

In the preceding section, we showed that every Lebesgue-integrable function f leads to an absolutely continuous function F via the indefinite integral. The following theorem shows that, up to an additive constant, there are no other AC functions; the key tools in the proof are the Jordan decomposition (5.1.14) and the Radon-Nikodym theorem proved in the preceding chapter (4.8.11).

5.2.1. Theorem. *The following conditions on a function $F : [a, b] \rightarrow \mathbb{R}$ are equivalent:*

(a) F is absolutely continuous;

¹³ E. Hewitt and K. Stromberg, *op. cit.*, p. 288, (18.25).

¹⁴ Cf. the author, "Lebesgue's 'Fundamental theorem of calculus' revisited" [*Paul Halmos: Celebrating 50 years of mathematics*, Springer-Verlag, New York, 1991, pp. 265–285], p. 284, Proposition. For an elementary proof of "AC \Rightarrow negligent", see *First course*, p. 206, Proposition 11.2.11.

(b) there exists a Lebesgue-integrable function $f : [a, b] \rightarrow \mathbb{R}$ such that

$$F(x) = F(a) + \int_a^x f \, d\lambda \quad \text{for all } x \in [a, b].$$

The function f is determined essentially uniquely by F , in the following sense: a Lebesgue-integrable function g on $[a, b]$ represents F as in (b) if and only if $g = f$ a.e. In particular, f can be taken to be a Borel function.

Proof. (b) \Rightarrow (a): This is immediate from 5.1.13.

(a) \Rightarrow (b): By the Jordan decomposition, it suffices to consider the case that F is increasing (5.1.14).

Consider first the case that F is *strictly* increasing. Writing $I = [a, b]$ and $J = F(I) = [F(a), F(b)]$, we have a homeomorphism $F : I \rightarrow J$. It follows that the correspondence $E \mapsto F(E)$ is a bijection between the Borel sets of I and those of J that preserves the set-theoretic operations. Let \mathcal{S} be the σ -algebra of Borel sets of I and define $\nu : \mathcal{S} \rightarrow [0, +\infty]$ by the formula

$$\nu(E) = \lambda(F(E)) \quad \text{for all } E \in \mathcal{S};$$

clearly ν is a finite measure on \mathcal{S} , and $\nu \ll \lambda$ by 5.1.16. By the Radon-Nikodym theorem (4.8.11) there exists a function $f : [a, b] \rightarrow \mathbb{R}$, measurable with respect to \mathcal{S} (hence a Borel function) and integrable with respect to λ , such that

$$\nu(E) = \int_E f \, d\lambda \quad \text{for all } E \in \mathcal{S}.$$

In particular for $E = [a, x]$, $a \leq x \leq b$, we have $F(E) = [F(a), F(x)]$ and $\nu(E) = \lambda(F(E)) = F(x) - F(a)$, thus

$$F(x) - F(a) = \int_{[a, x]} f \, d\lambda \quad \text{for all } x \in [a, b].$$

Suppose now that $F : [a, b] \rightarrow \mathbb{R}$ is any increasing AC function. Applying the preceding case to the strictly increasing AC function $F_1(x) = F(x) + x$, there exists a λ -integrable Borel function $f_1 : [a, b] \rightarrow \mathbb{R}$ such that

$$F_1(x) - F_1(a) = \int_{[a, x]} f_1 \, d\lambda \quad \text{for all } x \in [a, b],$$

that is,

$$F(x) - F(a) = -(x - a) + \int_a^x f_1 \, d\lambda$$

for all $x \in [a, b]$, thus the function $f = f_1 - 1$ meets the requirements of the theorem.

Finally, if f and g are Lebesgue-integrable functions on $[a, b]$, each of which represents F as in (b), then

$$\int_{[a, x]} f \, d\lambda = F(x) - F(a) = \int_{[a, x]} g \, d\lambda$$

for all $x \in [a, b]$, therefore $f = g$ λ -a.e. (4.7.7). \diamond

Exercise

1. If λ is Lebesgue measure on $[a, b]$ and ν is a finite measure defined on the Borel sets of $[a, b]$, such that $\lambda(B) = 0 \Rightarrow \nu(B) = 0$, then, by the Radon-Nikodym theorem (4.8.11) there exists a Lebesgue-integrable Borel function $f : [a, b] \rightarrow \mathbb{R}$ such that $\nu(B) = \int_B f d\lambda$ for all Borel sets $B \subset [a, b]$. Deduce this (circularly, of course!) from 5.2.1.

{Hint: Define a function $F : [a, b] \rightarrow \mathbb{R}$ by the formula $F(x) = \nu([a, x])$ and argue that F is absolutely continuous.¹}

5.3. limsup, liminf of Functions; Dini Derivates

The main mission of this section is to define the ‘Dini derivates’ of a function $g : [a, b] \rightarrow \mathbb{R}$; these will play a key role in the proof that a function of bounded variation is differentiable almost everywhere. As usual, it is worthwhile to see things in a larger perspective, so the definition of derivate is preceded by a general discussion of the underlying techniques (useful in other situations as well).

In §1.16, the concept of limit of a *sequence* of extended real numbers was created by fusing two more general concepts, limit superior and limit inferior (applicable to all sequences, not just those that have a limit). In an analogous way, we are going to dissect the concept of limit of a *function*, defined on a subset of a metric space and taking values in the extended reals $\overline{\mathbb{R}}$, into limit superior and limit inferior.

The general setup is the same as that for limits of functions (3.5.1). We have a metric space (X, d) and a function f defined on a subset B of X . In §3.5, f was allowed to take values in any metric space; here, we require f to have values in $\overline{\mathbb{R}}$, thus $f : B \rightarrow \overline{\mathbb{R}}$. We are interested in the behavior of $f(x)$ as x approaches a point $c \in X$, so c will at least have to be approximable by points of B , that is, $c \in \overline{B}$. We also want the option of restricting the *way* in which x approaches c (for example, if $X = \mathbb{R}$ we might want to require x to approach c from the left or from the right); in other words, we may want to specify a subset A of B and require x to approach c while remaining in A , so we will want c to be adherent to A . Thus, the framework for the discussion is as follows:

(X, d) is a metric space,

$$A \subset B \subset X, \quad f : B \rightarrow \overline{\mathbb{R}}, \quad c \in \overline{A}.$$

¹ Cf. the author, *Measure and integration* [Chelsea, New York, 1970], p. 149, §43, Theorem 1, or P.R. Halmos, *Measure theory* [Springer, 1974], p. 181, Theorem 43.D.

Schematically,

$$\begin{array}{ccc} & X & \\ & \cup & \\ & B & \xrightarrow{f} \overline{\mathbb{R}} \\ & \cup & \\ c \in \overline{A} & \supset & A \end{array}$$

5.3.1. *Definition.* With the preceding notations, for each neighborhood V of c in X , we have $V \cap A \neq \emptyset$; we write

$$\beta_V = \sup\{f(x) : x \in V \cap A\}$$

and define β to be the infimum of the β_V as V varies over the set of all neighborhoods of c :

$$\beta = \inf\{\beta_V : V \text{ a neighborhood of } c\}.$$

The extended real number β is called the **limit superior** of $f(x)$ as x approaches c through values in A , written

$$\beta = \limsup_{x \rightarrow c, x \in A} f(x) = \inf_V \left(\sup_{x \in V \cap A} f(x) \right),$$

where V runs over the set of all neighborhoods of c in X . Similarly, writing

$$\begin{aligned} \gamma_V &= \inf\{f(x) : x \in V \cap A\}, \\ \gamma &= \sup\{\gamma_V : V \text{ a neighborhood of } c\}, \end{aligned}$$

we call γ the **limit inferior** of $f(x)$ as x approaches c through values in A , written

$$\gamma = \liminf_{x \rightarrow c, x \in A} f(x) = \sup_V \left(\inf_{x \in V \cap A} f(x) \right).$$

If V and W are neighborhoods of c such that $V \subset W$, then $\beta_V \leq \beta_W$ and $\gamma_V \geq \gamma_W$. Adapting the notations of §1.16, we write

$$\beta_V \downarrow \beta \text{ and } \gamma_V \uparrow \gamma \text{ as } V \downarrow c.$$

A special case: If $A = B = X$, we write simply

$$\begin{aligned} \beta &= \limsup_{x \rightarrow c} f(x), \\ \gamma &= \liminf_{x \rightarrow c} f(x). \end{aligned}$$

That's a lot of machinery, but what is going on is very simple: as we shall see in the next theorem, β is the largest element of $\overline{\mathbb{R}}$ that is the limit of $f(x_n)$ for some sequence (x_n) in A with $x_n \rightarrow c$ (and γ is the smallest).

5.3.2. Lemma. If (V_n) is a sequence of neighborhoods of c such that $\text{diam } V_n \rightarrow 0$ (cf. 3.2.20), then $\beta_{V_n} \rightarrow \beta$ and $\gamma_{V_n} \rightarrow \gamma$ in $\overline{\mathbb{R}}$.

Proof. If $\beta = +\infty$ then $\beta_V = +\infty$ for all V , so the assertion about the β 's is trivial. Suppose $\beta < +\infty$. Assuming $r > \beta$, we have to show that $\beta \leq \beta_{V_n} < r$ ultimately. Choose a neighborhood V of c such that $\beta \leq \beta_V < r$ (possible because β is defined as a greatest lower bound). Since c is interior to V and $\text{diam } V_n \rightarrow 0$, $V_n \subset V$ ultimately, say for $n \geq N$; then $\beta \leq \beta_{V_n} \leq \beta_V < r$ for all $n \geq N$. This shows that $\beta_{V_n} \rightarrow \beta$; the proof that $\gamma_{V_n} \rightarrow \gamma$ is similar. \diamond

5.3.3. Theorem. With notations as above, let S be the set of all $\alpha \in \overline{\mathbb{R}}$ such that $f(x_n) \rightarrow \alpha$ for some sequence $x_n \in A$ with $x_n \rightarrow c$. Then β and γ belong to S , β is the largest element of S , and γ is the smallest; thus

$$\{\gamma, \beta\} \subset S \subset [\gamma, \beta],$$

where $[\gamma, \beta] = \{\alpha \in \overline{\mathbb{R}} : \gamma \leq \alpha \leq \beta\}$. In particular, $\gamma \leq \beta$, that is,

$$\liminf_{x \rightarrow c, x \in A} f(x) \leq \limsup_{x \rightarrow c, x \in A} f(x).$$

Proof. So to speak, the assertion is that β is the largest (extended real) number that $f(x)$ can be made to approach as x approaches c through points of A , and γ is the smallest.

Let $V_n = U_{1/n}(c) = \{x : d(x, c) < 1/n\}$; by the lemma, $\beta_{V_n} \rightarrow \beta$ and $\gamma_{V_n} \rightarrow \gamma$. Let us show, for example, that $\beta \in S$ (the proof that $\gamma \in S$ is similar).

case 1: $\beta = +\infty$.

Then $\beta_V = +\infty$ for every neighborhood V of c . In particular, $\beta_{V_n} = +\infty$, so there is a point $x_n \in V_n \cap A$ such that $f(x_n) > n$. Clearly $x_n \rightarrow c$ and $f(x_n) \rightarrow +\infty = \beta$, thus $\beta = +\infty \in S$.

case 2: $\beta = -\infty$.

For each n , choose a point $x_n \in V_n \cap A$; then $x_n \rightarrow c$ and $f(x_n) \leq \beta_{V_n} \rightarrow -\infty$, whence it is clear that $f(x_n) \rightarrow -\infty$, so that $\beta = -\infty \in S$.

case 3: $\beta \in \mathbb{R}$.

Then $\beta_{V_n} \in \mathbb{R}$ ultimately, say for $n \geq N$. For $1 \leq n < N$, choose any $x_n \in V_n \cap A$; for $n \geq N$ choose $x_n \in V_n \cap A$ so that

$$\beta_{V_n} - 1/n < f(x_n) \leq \beta_{V_n}$$

(possible because β_{V_n} is a least upper bound). Then $x_n \rightarrow c$ and $f(x_n) \rightarrow \beta$, thus $\beta \in S$.

We now show that β is the largest element of S (the proof that γ is the smallest is similar). Assuming $\alpha \in S$, we have to show that $\alpha \leq \beta$. This is trivial if $\beta = +\infty$ or if $\alpha = -\infty$; thus we can suppose that

$-\infty < \alpha$ and $\beta < +\infty$. We know that $f(x_n) \rightarrow \alpha$ for a sequence (x_n) in A such that $x_n \rightarrow c$. Since $\alpha > -\infty$, $f(x_n) > -\infty$ ultimately.

Fix $k \geq N$. Since $x_n \rightarrow c$ we have $x_n \in V_k$ ultimately, so

$$-\infty < f(x_n) \leq \beta_{V_k} < +\infty$$

for all sufficiently large n ; it follows that $\alpha \in \mathbb{R}$ and, passing to the limit as $n \rightarrow \infty$, we have $-\infty < \alpha \leq \beta_{V_k}$. Since $k \geq N$ is arbitrary, $\alpha \leq \beta$ by the lemma. \diamond

As noted in Example 3.3.17, $\overline{\mathbb{R}}$ can be equipped with a metric d' compatible with the concept of sequential convergence introduced in §1.16, that is, $\alpha_n \rightarrow \alpha$ in the sense of 1.16.8 if and only if $d'(\alpha_n, \alpha) \rightarrow 0$. The particular metric d' need not be specified; what counts is the property just mentioned. Viewing $\overline{\mathbb{R}}$ as a metric space, the present section fits into the general framework of limits in §3.5; the order structure of $\overline{\mathbb{R}}$ yields the following criterion for f to have a limit in the sense of §3.5:

5.3.4. Theorem. *With $f : B \rightarrow \overline{\mathbb{R}}$ as above, the following conditions are equivalent:*

(a) $\gamma = \beta$, that is,

$$\liminf_{x \rightarrow c, x \in A} f(x) = \limsup_{x \rightarrow c, x \in A} f(x);$$

(b) *the limit*

$$\lim_{x \rightarrow c, x \in A} f(x)$$

exists (in the sense of 3.5.1).

When this is the case, the three elements of $\overline{\mathbb{R}}$ in question are equal; expressed concisely,

$$\lim f = \liminf f = \limsup f$$

at the point c .

Proof. (b) \Rightarrow (a): Let

$$\alpha = \lim_{x \rightarrow c, x \in A} f(x);$$

this means (3.5.1) that if $x_n \in A$ and $x_n \rightarrow c$, then $f(x_n) \rightarrow \alpha$, thus the set S of 5.3.3 is a singleton, namely $S = \{\alpha\}$. By 5.3.3, $\{\gamma, \beta\} \subset \{\alpha\}$, so $\gamma = \beta = \alpha$.

(a) \Rightarrow (b): Write α for the common value of β and γ . By 5.3.3, $S = \{\alpha\}$. Assuming $x_n \in A$ and $x_n \rightarrow c$, we have to show that $f(x_n) \rightarrow \alpha$. At any rate, the sequence $(f(x_n))$ has a limit superior β' and a limit inferior γ' in the sense of §1.16; in view of the definition of limit given there (1.16.8), the problem is to show that $\beta' = \gamma' = \alpha$. By 1.16.11,

there exists a subsequence (x_{n_k}) of (x_n) such that $f(x_{n_k}) \rightarrow \beta'$; since $x_{n_k} \rightarrow c$, we have $\beta' \in S = \{\alpha\}$, thus $\beta' = \alpha$. Similarly, $\gamma' = \alpha$. \diamond

5.3.5. *Example.* Let $X = \mathbb{R}$ with the usual metric $d(x, y) = |x - y|$ and let $f : B \rightarrow \overline{\mathbb{R}}$, where $B \subset \mathbb{R}$. Let $c \in \mathbb{R}$ and suppose that

$$B \supset (c, c + r) \quad \text{for some } r > 0,$$

so that c is approachable from the right by points of B . Writing

$$A = B \cap (c, +\infty),$$

we have $c \in \overline{A}$ (in the event that $c \in B$, we have just masked it out), and the foregoing machinery is applicable. In the present situation, the limits superior and inferior of 5.3.1 are denoted

$$\limsup_{x \rightarrow c, x > c} f(x) \quad \text{and} \quad \liminf_{x \rightarrow c, x > c} f(x),$$

or, more concisely,

$$\limsup_{x \rightarrow c+} f(x) \quad \text{and} \quad \liminf_{x \rightarrow c+} f(x).$$

According to 5.3.4, these two numbers are equal if and only if

$$\lim_{x \rightarrow c, x \in A} f(x)$$

exists, that is (in the notations of 3.5.5), $f(c+)$ exists, in which case

$$f(c+) = \limsup_{x \rightarrow c+} f(x) = \liminf_{x \rightarrow c+} f(x).$$

One might reasonably write

$$f(c^+) = \limsup_{x \rightarrow c+} f(x),$$

$$f(c_+) = \liminf_{x \rightarrow c+} f(x),$$

so that

$$f(c+) \text{ exists} \Leftrightarrow f(c_+) = f(c^+),$$

in which case $f(c+) = f(c_+) = f(c^+)$; this would place too much strain on the notation and we shall not do so. (A minor variation on this idea is in common use for the 'derivates' to be defined shortly, but in that context the signs are positioned so that it is easier for the eye to sort them out.)

Similarly, if $B \supset (c - r, c)$ for some $r > 0$, the numbers

$$\limsup_{x \rightarrow c-} f(x) \quad \text{and} \quad \liminf_{x \rightarrow c-} f(x)$$

are defined in the expected way: we set $A = B \cap (-\infty, c)$ and again apply the machinery of 5.3.1. Thus, these two numbers are equal if and only if

$f(c-)$ exists in the sense of 3.5.5, in which case all three numbers are equal.

Taking into account the criterion of 3.5.6 for the existence of a (two-sided) limit, we have:

5.3.6. Theorem. Let $B \subset \mathbb{R}$, $f : B \rightarrow \overline{\mathbb{R}}$, $c \in \mathbb{R}$, and suppose that $B \supset (c-r, c) \cup (c, c+r)$ for some $r > 0$. In order that

$$\lim_{x \rightarrow c, x \neq c} f(x)$$

exist (in the sense of 3.5.5), it is necessary and sufficient that the four numbers

$$\begin{aligned} \limsup_{x \rightarrow c+} f(x), \quad \liminf_{x \rightarrow c+} f(x), \\ \limsup_{x \rightarrow c-} f(x), \quad \liminf_{x \rightarrow c-} f(x), \end{aligned}$$

be equal, in which case all five numbers are equal.

5.3.7. Definition. Let $g : [a, b] \rightarrow \mathbb{R}$, $a < b$, and let $c \in [a, b]$. Write $B = [a, b] - \{c\}$ and define $f : B \rightarrow \overline{\mathbb{R}}$ by the formula

$$f(x) = \frac{g(x) - g(c)}{x - c}.$$

Of course the values of f are in \mathbb{R} , but we are being consistent with the foregoing notations; some of the numbers we are about to associate with f may be infinite.

If $c \in [a, b)$ then c is approachable from the right by $x \in B$ and we define

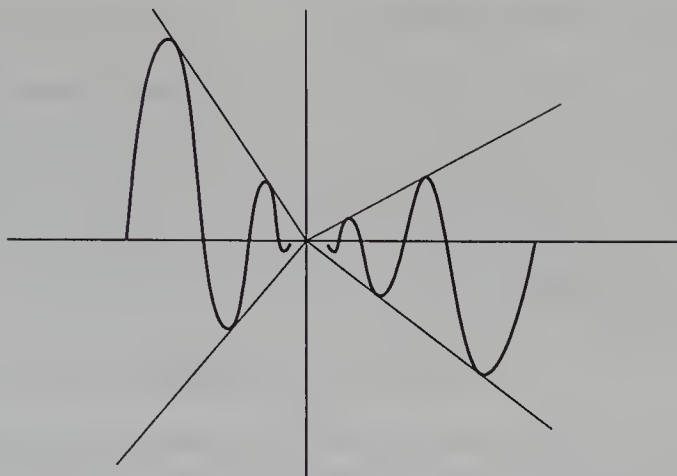
$$\begin{aligned} (D^+g)(c) &= \limsup_{x \rightarrow c+} f(x) = \limsup_{x \rightarrow c+} \frac{g(x) - g(c)}{x - c}, \\ (D_+g)(c) &= \liminf_{x \rightarrow c+} f(x) = \liminf_{x \rightarrow c+} \frac{g(x) - g(c)}{x - c}. \end{aligned}$$

Similarly, if $c \in (a, b]$ we define

$$\begin{aligned} (D^-g)(c) &= \limsup_{x \rightarrow c-} f(x) = \limsup_{x \rightarrow c-} \frac{g(x) - g(c)}{x - c}, \\ (D_-g)(c) &= \liminf_{x \rightarrow c-} f(x) = \liminf_{x \rightarrow c-} \frac{g(x) - g(c)}{x - c}. \end{aligned}$$

These four numbers are called the **Dini derivatives** of g at c ; more precisely (for example), $(D_+g)(c)$ is the **lower right-hand derivate** of g at c .

5.3.8. Example. When $a < c < b$, all four derivatives are defined; they may all be different, as in the following variation on $x \sin(1/x)$:



5.3.9. *Example.* When g is increasing, all of its derivatives are ≥ 0 . The preoccupation of the next section is with the reverse implication: under certain conditions, monotonicity can be inferred from the sign of the derivatives.

Definition 5.3.7 and Theorem 5.3.6 allow us to refine the discussion of derivatives in §3.5:

5.3.10. *Theorem.* Let $g : [a, b] \rightarrow \mathbb{R}$, $a < b$.

(i) In order that g be **right-differentiable** at $c \in [a, b)$, it is necessary and sufficient that its upper and lower right-hand derivatives at c be **finite and equal**, in which case

$$g'_r(c) = (D^+g)(c) = (D_+g)(c).$$

(ii) Similarly, g is **left-differentiable** at $c \in (a, b]$ if and only if its upper and lower left-hand derivatives at c are **finite and equal**, in which case

$$g'_l(c) = (D^-g)(c) = (D_-g)(c).$$

(iii) g is **differentiable** at c if and only if all four of its derivatives at c are **finite and equal**, in which case $g'(c)$ is the common value of the derivatives.

Exercise

1. If $f : X \rightarrow \overline{\mathbb{R}}$ and $c \in X$, then

$$\liminf_{x \rightarrow c} f(x) \leq f(c) \leq \limsup_{x \rightarrow c} f(x).$$

5.4. Criteria for Monotonicity

The (relatively) elementary theorem in this circle of ideas:

If $f : [a, b] \rightarrow \mathbb{R}$ is continuous and f is differentiable on (a, b) , then f is increasing on $[a, b]$ if and only if $f'(x) \geq 0$ for all $x \in (a, b)$.

{The implication $\ll f' \geq 0 \Rightarrow f \text{ increasing} \gg$ entails the mean value theorem¹; the reverse implication follows trivially from the definitions.} Loosely speaking,

$$f' \geq 0 \Leftrightarrow f \text{ increasing}$$

provided that f is ‘sufficiently smooth’. When one tries to relax (or merely change) the conditions of ‘smoothness’, the game gets very delicate. Here are three landmark theorems, in increasing order of difficulty:

Theorem A. *If $f : [a, b] \rightarrow \mathbb{R}$ is absolutely continuous and one of its four derivatives is ≥ 0 a.e., then f is increasing.*

{Compared to the “elementary theorem” mentioned above, the function f in Theorem A is required to be ‘more continuous’ and permitted to be ‘less differentiable’.

Theorem B. *If $f : [a, b] \rightarrow \mathbb{R}$ is continuous and one of its four derivatives is ≥ 0 for all but countably many points of (a, b) , then f is increasing.*

{Theorem B is a true generalization of the “elementary theorem”; the continuity hypothesis is the same, but the differentiability hypothesis is weaker.}

Theorem C. *If $f : [a, b] \rightarrow \mathbb{R}$ is increasing, then $f'(x)$ exists and is ≥ 0 for almost every $x \in (a, b)$.*

{One of the high-water marks of the Lebesgue theory, Theorem C is a striking extension of the implication “ \Leftarrow ” of the “elementary theorem”; continuity goes out the window², but differentiability slips back in through the screen door.}

Theorems A and B are proved in the present section, Theorem C in §12 (5.12.7 below). The proofs to be given are patterned after E. J. McShane’s elegant exposition.³ By way of orientation, we begin with an easy theorem about monotonicity.

5.4.1. Theorem. *If I is an interval and $f : I \rightarrow \mathbb{R}$ is monotone, then the set of discontinuities of f is countable.*

¹ *First course*, p. 131, Corollary 8.5.3.

² However, it is elementary that a monotone function has at most countably many discontinuities (5.4.1 below).

³ E. J. McShane, *Integration*, Princeton University Press, Princeton, N.J., 1944; an authentic, user-friendly textbook, not the forbidding treatise it may seem at first glance.

Proof. We can suppose that f is increasing (if not, switch to $-f$) and that $I = [a, b]$ (every interval is the union of an increasing sequence of closed intervals). For each $x \in (a, b)$, f has a left limit $f(x-)$ and a right limit $f(x+)$, with $f(x-) \leq f(x) \leq f(x+)$; for f to be discontinuous at x , it is necessary and sufficient that $f(x+) - f(x-) > 0$. The set of discontinuities of f in (a, b) is the union of the sets

$$S_n = \{x \in (a, b) : f(x+) - f(x-) > 1/n\}$$

($n = 1, 2, 3, \dots$), so it suffices to show that each S_n is finite.

To simplify the notation, let $r > 0$ and let

$$S = \{x \in (a, b) : f(x+) - f(x-) > r\};$$

we need only show that S is finite. Suppose x_1, \dots, x_n are points of S with $x_1 < x_2 < \dots < x_n$. Choose points c_1, \dots, c_n, c_{n+1} in (a, b) such that $c_1 < x_1$, $x_{i-1} < c_i < x_i$ for $i = 2, \dots, n$, and $x_n < c_{n+1}$. For $i = 1, \dots, n$ we have $c_i < x_i < c_{i+1}$, whence

$$f(c_i) \leq f(x_{i-}) \leq f(x_i+) \leq f(c_{i+1}),$$

so that

$$f(x_i+) - f(x_{i-}) \leq f(c_{i+1}) - f(c_i);$$

summing over i and telescoping the sum on the right side, we have

$$\sum_{i=1}^n [f(x_i+) - f(x_{i-})] \leq f(c_{n+1}) - f(c_1) \leq f(b) - f(a).$$

whence $f(b) - f(a) > nr$. This shows that n is bounded, so S cannot be infinite. \diamond

The following theorem is somewhat more general than our immediate needs in this section (a proof of Theorems A and B of the introduction) but one of its corollaries (5.4.7) is needed for the proof of Theorem C (a.e. differentiability of monotone functions) in §5.12.⁴

5.4.2. Theorem. Suppose $f : [a, b] \rightarrow \mathbb{R}$ satisfies the conditions

(i) $f(x) \leq \liminf_{t \rightarrow x+} f(t)$ for all $x \in [a, b)$,

(ii) $\limsup_{t \rightarrow x-} f(t) \leq f(x)$ for all $x \in (a, b]$,

and let $N = \{x \in (a, b) : (D^+ f)(x) \leq 0\}$.

If $f(N)$ has empty interior, then f is increasing; if, in addition, N has empty interior, then f is strictly increasing.

⁴ It seems paradoxical that a theorem whose *conclusion* is monotonicity should figure in the proof of a theorem *about* monotone functions, which is to say that McShane's proof of Theorem C is a *very* cunning proof.

Proof. {Note that (i) and (ii) are *necessary* conditions for f to be increasing. If f is continuous on $[a, b]$ then (i) and (ii) hold trivially, with equality. To say that N has empty interior means that every point of N is a boundary point; equivalently, $[a, b] - N$ is dense in $[a, b]$, in other words, the set of all points x with $(D^+ f)(x) > 0$ is dense in $[a, b]$.}

Assuming f satisfies (i)–(ii) and $f(N)$ has empty interior, we must show that f is increasing. Suppose to the contrary that there exists a pair of points c, d with $a \leq c < d \leq b$ and $f(c) > f(d)$. Since $f(N)$ has empty interior, it cannot contain the interval $(f(d), f(c))$; choose a point $k \in (f(d), f(c))$ that does not belong to $f(N)$, that is,

$$f(d) < k < f(c) \quad \text{and} \quad (\forall x \in N) f(x) \neq k.$$

Consider the set

$$S = \{x \in [c, d] : f(x) \geq k\};$$

for example, $c \in S$. Defining $s = \sup S$, we have $c \leq s \leq d$. We will show that each of the alternatives $f(s) > k$, $f(s) < k$, $f(s) = k$ leads to a contradiction.

(1) If $f(s) > k$ then $s \neq d$ (because $f(d) < k$), therefore $s < d$. For every $x \in (s, d)$ we have

$$x > s = \sup S,$$

hence $x \notin S$, therefore $f(x) < k$ by the definition of S . Thus

$$\sup_{x \in (s, d)} f(x) \leq k;$$

since (s, d) is a deleted right neighborhood of s , it follows from the definition of limit superior (as an inf of sups) that

$$\limsup_{x \rightarrow s+} f(x) \leq \sup_{x \in (s, d)} f(x) \leq k < f(s),$$

therefore also

$$\liminf_{x \rightarrow s+} f(x) < f(s),$$

which contradicts hypothesis (i).

(2) If $f(s) < k$ then $s \neq c$ (because $f(c) > k$), therefore $c < s$. For each r such that $c \leq r < s = \sup S$, there exists a point $t \in S$ with $r < t \leq s$, and we have $f(t) \geq k$ by the definition of S ; necessarily $t \neq s$ (because $f(s) < k$ by supposition), thus $r < t < s$, that is, $t \in (r, s)$. It follows that

$$\sup_{x \in (r, s)} f(x) \geq k;$$

taking the infimum of the left sides over all possible r , we have

$$\limsup_{x \rightarrow s-} f(x) \geq k > f(s),$$

which contradicts hypothesis (ii).

(3) If $f(s) = k$ then $f(d) < f(s) < f(c)$; in particular, $s \neq d$, $s \neq c$, so $c < s < d$. For every $x \in (s, d)$ we have $x \in [c, d]$ but $x \notin S$ (because $x > s = \sup S$), therefore $f(x) < k$ by the definition of S , that is, $f(x) < f(s)$; thus

$$(*) \quad \frac{f(x) - f(s)}{x - s} < 0$$

for all $x \in (s, d)$. Then

$$\sup_{x \in (s, d)} \left(\frac{f(x) - f(s)}{x - s} \right) \leq 0;$$

since (s, d) is a deleted right neighborhood of s , it follows from the definition of $(D^+f)(s)$ as a limsup (an inf of sups) that $(D^+f)(s) \leq 0$. Thus $s \in N$, and $k = f(s) \in f(N)$ contradicts the choice of k .

The final assertion of the theorem is elementary: an increasing function $f : [a, b] \rightarrow \mathbb{R}$ for which the set N has empty interior must be strictly increasing. For, the alternative is that there exist points c and d with $a \leq c < d \leq b$ and $f(c) = f(d)$; then f is constant on $[c, d]$, $(D^+f)(x) = 0$ for all $x \in (c, d)$, and N contains the nonempty open set (c, d) , a contradiction. \diamond

The particular derivate in the above theorem is not critical:

5.4.3. Corollary. Suppose $f : [a, b] \rightarrow \mathbb{R}$ satisfies the conditions (i) and (ii) of Theorem 5.4.2, D is any one of the derivate operations D^+ , D_+ , D^- , D_- , and $N = \{x \in (a, b) : (Df)(x) \leq 0\}$.

If $f(N)$ has empty interior, then f is increasing; if, in addition, N has empty interior, then f is strictly increasing.

Proof. For $D = D^+$ this is the theorem.

Suppose $D = D_+$. Writing

$$N^+ = \{x \in (a, b) : (D^+f)(x) \leq 0\},$$

it follows from $D^+f \geq D_+f$ that $N^+ \subset N$, whence $f(N^+) \subset f(N)$. If $f(N)$ has empty interior, the same is true of $f(N^+)$, therefore f is increasing by the first case; if, moreover, N has empty interior, then the same is true of N^+ , therefore f is strictly increasing.

Suppose $D = D^-$. The strategy is to construct a function f^* for which the range of D^+f^* equals the range of D^-f , then infer the corollary for f by applying the first case to f^* . For $x \in [a, b]$, write

$$x^* = a + b - x;$$

the mapping $x \mapsto x^*$ is an order-reversing bijection $[a, b] \rightarrow [a, b]$ such that

$$x^{**} = x, \quad a^* = b, \quad b^* = a \quad \text{and} \quad x^* - y^* = y - x$$

for all x, y in $[a, b]$. Define $f^* : [a, b] \rightarrow \mathbb{R}$ by

$$f^*(x) = -f(x^*).$$

Since $x \mapsto x^*$ and $y \mapsto -y$ are both order-reversing, f will be increasing (strictly increasing) if and only if f^* is increasing (strictly increasing).

Multiplying through the inequality in (i) by -1 , we have

$$-f(x) \geq -\liminf_{t \rightarrow x+} f(t) = \limsup_{t \rightarrow x+} [-f(t)]$$

for all $x \in [a, b]$, in other words,

$$f^*(x^*) \geq \limsup_{t^* \rightarrow x^*-} f^*(t^*)$$

for all $x^* \in (b^*, a^*] = (a, b]$, thus f^* satisfies condition (ii) of the theorem. Similarly, since f satisfies (ii), it follows that f^* satisfies (i); explicitly, for all $x \in (a, b]$,

$$-f(x) \leq -\limsup_{t \rightarrow x-} f(t) = \liminf_{t \rightarrow x-} [-f(t)],$$

that is,

$$f^*(x^*) \leq \liminf_{t^* \rightarrow x^*+} f^*(t^*)$$

for all $x^* \in [b^*, a^*) = [a, b]$.

Thus, f^* satisfies conditions (i) and (ii). For all $x \in (a, b)$ we have $(D^+ f^*)(x) = (D^- f)(x^*)$; for,

$$\begin{aligned} \limsup_{t \rightarrow x+} \frac{f^*(t) - f^*(x)}{t - x} &= \limsup_{t \rightarrow x+} \frac{-f(t^*) + f(x^*)}{x^* - t^*} \\ &= \limsup_{t^* \rightarrow x^*-} \frac{f(t^*) - f(x^*)}{t^* - x^*}. \end{aligned}$$

Let

$$N^* = \{x \in (a, b) : (D^+ f^*)(x) \leq 0\};$$

since $(D^+ f^*)(x) = (D^- f)(x^*)$, we have

$$\begin{aligned} N^* &= \{x \in (a, b) : (D^- f)(x^*) \leq 0\} \\ &= \{x \in (a, b) : x^* \in N\}, \end{aligned}$$

that is, $x \in N^* \Leftrightarrow x^* \in N$. Then

$$\begin{aligned} f^*(N^*) &= \{f^*(x) : x \in N^*\} \\ &= \{-f(x^*) : x^* \in N\} = -f(N). \end{aligned}$$

If $f(N)$ has empty interior, then so does $f^*(N^*) = -f(N)$, therefore f^* (hence also f) is increasing by the theorem. Suppose, in addition,

that N has empty interior; since $x \mapsto x^*$ is a homeomorphism of $[a, b]$ that transforms N into N^* , it follows that N^* also has empty interior, therefore f^* (hence also f) is strictly increasing.

The case that $D = D_-$ follows from the preceding case by the argument used for the case that $D = D_+$. \diamond

It will be shown later in the chapter that an absolutely continuous function is differentiable almost everywhere (5.9.4); much more accessible is the fact that the behavior of such a function is controlled by the sign of its derivatives:

5.4.4. Corollary. *Suppose $f : [a, b] \rightarrow \mathbb{R}$ is absolutely continuous and let D be one of D^+, D_+, D^-, D_- . If $(Df)(x) > 0$ a.e. in (a, b) , then f is strictly increasing on $[a, b]$.*

Proof. Let $N = \{x \in (a, b) : (Df)(x) \leq 0\}$. By assumption, N is negligible; since f is AC, $f(N)$ is also negligible (5.1.16). Negligible sets obviously have empty interior, so f is strictly increasing by the preceding corollary. \diamond

The next corollary disposes of Theorem A of the introduction:

5.4.5. Corollary. *Suppose $f : [a, b] \rightarrow \mathbb{R}$ is absolutely continuous and let D be one of D^+, D_+, D^-, D_- . Then:*

- (1) $Df \geq 0$ a.e. $\Rightarrow f$ increasing;
- (2) $Df = 0$ a.e. $\Rightarrow f$ constant.

Proof. (1) For each $r > 0$, consider the function $f_r : [a, b] \rightarrow \mathbb{R}$ defined by

$$f_r(x) = f(x) + rx.$$

Then f_r is AC (it is the sum of two AC functions), and

$$(Df_r)(x) = (Df)(x) + r \geq 0 + r > 0$$

for almost all x , therefore f_r is strictly increasing by the preceding corollary. Thus, if $a \leq c < d \leq b$ then, for every $r > 0$, $f_r(c) < f_r(d)$, that is,

$$f(c) + rc < f(d) + rd,$$

and $f(c) \leq f(d)$ results on letting $r \rightarrow 0$.

(2) Suppose, for example, that $D^+f = 0$ a.e. By (1), f is increasing. But

$$D_+(-f) = -D^+f = 0 \text{ a.e.},$$

so $-f$ is also increasing. Thus f is both increasing and decreasing, therefore constant. \diamond

On the way to Theorem B,

5.4.6. Corollary. *Suppose $f : [a, b] \rightarrow \mathbb{R}$ satisfies conditions (i) and (ii) of Theorem 5.4.2, and let D be one of D^+, D_+, D^-, D_- .*

If $(Df)(x) > 0$ for all but countably many x in (a, b) , then f is strictly increasing on $[a, b]$.

Proof. By assumption, the set

$$N = \{x \in (a, b) : (Df)(x) \leq 0\}$$

is countable, therefore so is $f(N)$, thus both N and $f(N)$ have empty interior. \diamond

5.4.7. Corollary. Suppose $f : [a, b] \rightarrow \mathbb{R}$ satisfies conditions (i) and (ii) of Theorem 5.4.2, and let D be one of D^+, D_+, D^-, D_- .

(1) If $(Df)(x) \geq 0$ for all but countably many x in (a, b) , then f is increasing.

(2) If $(Df)(x) = 0$ for all but countably many x in (a, b) , then f is constant.

Proof. (1) As in the proof of 5.4.5, write $f_r(x) = f(x) + rx$ for $r > 0$. It is clear that each f_r satisfies the conditions (i) and (ii) (because rx is continuous; cf. 5.3.3). Moreover,

$$(Df_r)(x) = (Df)(x) + r \geq 0 + r > 0$$

for all but countably many x , so every f_r is (strictly) increasing by the preceding corollary, therefore f is increasing by the argument in 5.4.5.

(2) In the proof of (2) of 5.4.5, replace “a.e.” by “at all but countably many points”. \diamond

The next corollary includes Theorem B of the introduction:

5.4.8. Corollary. Suppose $f : [a, b] \rightarrow \mathbb{R}$ is continuous and let D be one of D^+, D_+, D^-, D_- .

(1) If $Df \geq 0$ at all but countably many points, then f is increasing.

(2) If $Df = 0$ at all but countably many points, then f is constant.

Proof. Conditions (i) and (ii) of Theorem 5.4.2 are trivially verified by f , so the present corollary is immediate from 5.4.7. \diamond

Corollary 5.4.7 also yields a criterion for a monotone function to be strictly monotone:

5.4.9. Corollary. Let $f : [a, b] \rightarrow \mathbb{R}$ be increasing (so that all four derivatives of f are ≥ 0), let D be one of D^+, D_+, D^-, D_- and let

$$A = \{x \in (a, b) : (Df)(x) > 0\}.$$

Then f is strictly increasing on $[a, b]$ if and only if A is dense in $[a, b]$.

Proof. Arguing contrapositively, let us show that (assuming f increasing)

$$f \text{ not strictly increasing} \Leftrightarrow A \text{ not dense in } [a, b].$$

\Rightarrow : Assuming f increasing but not strictly, the argument at the end of the proof of Theorem 5.4.2 shows that the set $N = (a, b) - A$ has nonempty interior, consequently A is not dense.

\Leftarrow : Let (c, d) be an open subinterval of $[a, b]$ that contains no point of A . Then $(Df)(x) = 0$ for all $x \in (c, d)$. The restricted function $g = f|_{[c, d]}$ is increasing, therefore satisfies (i), (ii) of 5.4.2; moreover, $(Dg)(x) = 0$ for all x in (c, d) , so g is constant on $[c, d]$ by (2) of Corollary 5.4.7. Thus $f(c) = f(d)$ and f is not strictly increasing. \diamond

Exercises

1. Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous and suppose N is a subset of $[a, b]$ such that, at every $x \in (a, b) - N$, f is differentiable and $f'(x) > 0$.

(i) If $f(N)$ has empty interior, then f is increasing.

(ii) If both $f(N)$ and N have empty interior then f is strictly increasing.

(iii) Infer that if f is absolutely continuous and N is negligible, then f is strictly increasing.

{Hint: (i), (ii) Simplify the proof of Theorem 5.4.2.}

2. In (ii) of Exercise 1, the assumption that N has empty interior cannot be omitted.

{Hint: Let $f : [-1, 1] \rightarrow \mathbb{R}$ be the function such that $f(x) = 0$ on $[-1, 0]$ and $f(x) = x$ on $(0, 1]$, and let $N = [-1, 0]$.}

5.5. Semicontinuity

In the next section, we prove a theorem on the approximation of Lebesgue-integrable functions by ‘semicontinuous’ ones (for application in the proof of Lebesgue’s “Fundamental theorem of calculus”); the present section lays the necessary technical foundation. As the section heading suggests, the concept of continuity (for extended-real-valued functions) is to be dissected into two more general concepts, upper semicontinuity and lower semicontinuity. Except that the functions are to take values in $\overline{\mathbb{R}}$, the setting is purely topological; throughout the section, X is a topological space¹ (specialized to $[a, b]$ from Definition 5.5.22 onward).

5.5.1. *Definition.* Let c be a point of X . A function $f : X \rightarrow \overline{\mathbb{R}}$ is said to be **lower semicontinuous** (l.s.c.) at c if, for every real number r such that $f(c) > r$, the set $\{x \in X : f(x) > r\}$ is a neighborhood of c . {Informally, every inequality $f(c) > r$ persists on a neighborhood of c .}

¹ Metric spaces are sufficient for our applications.

5.5.2. *Example.* Let $A \subset X$. The characteristic function $f = \varphi_A$ of A is l.s.c. at every point of $X - A$. If $c \in A$, then f is l.s.c. at $c \Leftrightarrow c$ is interior to A .

{Proof: If $c \in X - A$ then $f(c) = 0$; for every $r < f(c) = 0$, the set $\{x : f(x) > r\} = X$ is a neighborhood of c . Suppose $c \in A$, so that $f(c) = 1$; if $r < f(c) = 1$ then the set $\{x : f(x) > r\}$ is either A or X according as $r \geq 0$ or $r < 0$, guaranteed to be a neighborhood of c if and only if A is a neighborhood of c .}

5.5.3. *Remarks.* In the notations of Definition 1, if $f(c) = -\infty$ then f is l.s.c. at c 'by default' ($f(c) > r$ is impossible). If f is l.s.c. at c , then so is af ($a > 0$).

5.5.4. *Theorem.* Let $f_i : X \rightarrow \overline{\mathbb{R}}$ ($i \in I$) be a family of functions, $c \in X$, and $f = \sup f_i$ the upper envelope of the family, that is, $f(x) = \sup\{f_i(x) : i \in I\}$ for all $x \in X$.

If every f_i is lower semicontinuous at c , then so is f .

Proof. Assuming $f(c) > r \in \mathbb{R}$, we have to show that $\{x : f(x) > r\}$ is a neighborhood of c . Since

$$f(x) \leq r \Leftrightarrow f_i(x) \leq r \text{ for all } i \in I,$$

we have

$$f(x) > r \Leftrightarrow f_i(x) > r \text{ for some } i \in I,$$

thus

$$(*) \quad \{x : f(x) > r\} = \bigcup_{i \in I} \{x : f_i(x) > r\}.$$

Since c belongs to the left side of (*), it belongs to some term on the right, say $c \in \{x : f_j(x) > r\}$. In particular, $f_j(c) > r$; since f_j is l.s.c. at c , $\{x : f_j(x) > r\}$ is a neighborhood of c , therefore so is the left side of (*). \diamond

For finite families, the analogous result holds for lower envelopes:

5.5.5. *Theorem.* If each of the functions $f_i : X \rightarrow \overline{\mathbb{R}}$ ($i = 1, \dots, n$) is lower semicontinuous at $c \in X$, then so is their lower envelope $f = \inf(f_1, \dots, f_n)$.

Proof. Suppose $f(c) > r \in \mathbb{R}$. For $i = 1, \dots, n$, $f_i(c) \geq f(c) > r$, therefore $\{x : f_i(x) > r\}$ is a neighborhood of c by the lower semicontinuity of f_i at c ; thus

$$\{x : f(x) > r\} = \bigcap_{i=1}^n \{x : f_i(x) > r\}$$

is the intersection of finitely many neighborhoods of c . \diamond

It is time to drop the other shoe:

5.5.6. *Definition.* With notations as in Definition 5.5.1, f is said to be **upper semicontinuous** (u.s.c.) at c if, for every real number r such that $f(c) < r$, the set $\{x : f(x) < r\}$ is a neighborhood of c . {So to speak, every inequality $f(c) < r$ persists on a neighborhood of c .}

5.5.7. *Remarks.* (1) f is u.s.c. at $c \Leftrightarrow -f$ is l.s.c. at c . {Cf. 1.15.4, (i).}

(2) If $f(c) = +\infty$ then f is u.s.c. at c .

(3) Let $A \subset X$. The characteristic function $f = \varphi_A$ is u.s.c. at every point of A . If $c \in X - A$, then f is u.s.c. at $c \Leftrightarrow c$ is interior to $X - A$ (that is, exterior to A).

(4) If f is u.s.c. at c , then so is af ($a > 0$).

In view of the above remark (1), Theorems 5.5.4 and 5.5.5 yield the ‘dual’ statements:

5.5.8. *Theorem.* Let $f_i : X \rightarrow \overline{\mathbb{R}}$ ($i \in I$) be a family of functions and let $c \in X$.

(i) If every f_i is upper semicontinuous at c , then so is the lower envelope $\inf f_i$ of the family.

(ii) If $I = \{1, \dots, n\}$ and if every f_i is upper semicontinuous at c , then so is the upper envelope $\sup(f_1, \dots, f_n)$ of the family.

Upper and lower semicontinuity are a ‘dissection’ of continuity in the following sense:

5.5.9. *Theorem.* For a function $f : X \rightarrow \overline{\mathbb{R}}$ and a point $c \in X$, the following conditions are equivalent:

(a) f is continuous at c ;

(b) f is both lower and upper semicontinuous at c .

Proof. (a) \Rightarrow (b): We are assuming that f is continuous at c . Suppose first that $f(c) \in \mathbb{R}$. If $r < f(c) < s$ then the interval (r, s) is a neighborhood of $f(c)$, therefore the set

$$\{x : r < f(x) < s\} = f^{-1}((r, s))$$

is a neighborhood of c ; since

$$(*) \quad \{x : r < f(x) < s\} = \{x : f(x) > r\} \cap \{x : f(x) < s\},$$

it is clear that $\{x : f(x) > r\}$ and $\{x : f(x) < s\}$ are both neighborhoods of c . This shows that f is both l.s.c. and u.s.c. at c .

If $f(c) = +\infty$ then f is ‘automatically’ u.s.c. at c . For every $r \in \mathbb{R}$, we have $f(c) > r$ and $(r, +\infty]$ is a neighborhood of $+\infty = f(c)$ in $\overline{\mathbb{R}}$, therefore the set

$$\{x : f(x) > r\} = f^{-1}((r, +\infty])$$

is a neighborhood of c ; thus f is also l.s.c. at c .

Finally, if $f(c) = -\infty$ then f is l.s.c. at c . For every $r \in \mathbb{R}$, $f(c) < r$ and $[-\infty, r)$ is a neighborhood of $-\infty = f(c)$, therefore

$$\{x : f(x) < r\} = f^{-1}([-\infty, r))$$

is a neighborhood of c ; thus f is also u.s.c. at c .

(b) \Rightarrow (a): We are assuming that f is both l.s.c. and u.s.c. at c . Suppose first that $f(c) \in \mathbb{R}$. If $r < f(c) < s$ then (r, s) is a basic neighborhood of $f(c)$ in $\overline{\mathbb{R}}$; from (*) we see that $f^{-1}((r, s))$ is the intersection of two neighborhoods of c , hence is a neighborhood of c . This shows that f is continuous at c .

If $f(c) = +\infty$ then $f(c) > r$ for every $r \in \mathbb{R}$ and the sets $(r, +\infty]$ are basic neighborhoods of $+\infty = f(c)$; by assumption, the sets

$$f^{-1}((r, +\infty]) = \{x : f(x) > r\}$$

are neighborhoods of c , whence the continuity of f at c .

Similarly, if $f(c) = -\infty$ and $r \in \mathbb{R}$ then $f^{-1}([-\infty, r)) = \{x : f(x) < r\}$ is a neighborhood of c , therefore f is continuous at c . \diamond

5.5.10. *Definition.* A function $f : X \rightarrow \overline{\mathbb{R}}$ is said to be **lower semicontinuous** (l.s.c.) **on** X if it is l.s.c. at every point of X ; it is said to be **upper semicontinuous** (u.s.c.) **on** X if it is u.s.c. at every point of X .

5.5.11. *Example.* For a subset A of X ,

$$\varphi_A \text{ l.s.c.} \Leftrightarrow A \text{ open,}$$

$$\varphi_A \text{ u.s.c.} \Leftrightarrow A \text{ closed}$$

(see the remarks following Definitions 5.5.1 and 5.5.6).

5.5.12. *Theorem.* The following conditions on a function $f : X \rightarrow \overline{\mathbb{R}}$ are equivalent:

(a) f is lower semicontinuous on X ;

(b) for every real number r , $\{x : f(x) > r\}$ is an open set in X .

Proof. (a) \Rightarrow (b): Let $r \in \mathbb{R}$, $A = \{x : f(x) > r\}$. For every $c \in A$, A is a neighborhood of c (Definition 5.5.1), thus every point of A is an interior point.

(b) \Rightarrow (a): Let $c \in X$. For every real number $r < f(c)$, the set $\{x : f(x) > r\}$ is open, hence is a neighborhood of c , thus f is l.s.c. at c . \diamond

5.5.13. *Corollary.* A function $f : X \rightarrow \overline{\mathbb{R}}$ is upper semicontinuous on X if and only if, for every real number r , $\{x : f(x) < r\}$ is an open set.

5.5.14. *Theorem.* Let $f_i : X \rightarrow \overline{\mathbb{R}}$ ($i \in I$) be a family of functions.

(1) If every f_i is l.s.c. on X , then so is $\sup f_i$; if, moreover, I is finite, then $\inf f_i$ is also l.s.c. on X .

(2) If every f_i is u.s.c. on X , then so is $\inf f_i$; if, moreover, I is finite, then $\sup f_i$ is also u.s.c. on X .

Proof. Immediate from Theorems 5.5.4, 5.5.5, 5.5.8 and the definitions. \diamond

In §5.3, \liminf and \limsup were defined in the context of a metric space; inspection of Definition 5.3.1 shows that the concepts carry over verbatim to functions defined on a topological space. (However, the proofs of the results in §5.3 were based on sequential convergence in metric spaces, so we must be careful not to cite these results without revisiting the proofs.) The form of the definitions we need here are as follows:

5.5.15. *Definition.* Let $f : X \rightarrow \overline{\mathbb{R}}$ and let $c \in X$. For each neighborhood V of c in X , we define

$$\gamma_V = \inf_{x \in V} f(x), \quad \beta_V = \sup_{x \in V} f(x).$$

Letting V vary over the set of all neighborhoods of c in X , we define

$$\gamma = \sup\{\gamma_V : V \text{ a neighborhood of } c\}$$

and call it the **limit inferior** of f at c , written

$$\liminf_{x \rightarrow c} f(x) = \gamma = \sup_V \left(\inf_{x \in V} f(x) \right).$$

Similarly, we define

$$\beta = \inf\{\beta_V : V \text{ a neighborhood of } c\}$$

and call it the **limit superior** of f at c , written

$$\limsup_{x \rightarrow c} f(x) = \beta = \inf_V \left(\sup_{x \in V} f(x) \right).$$

As noted following 5.3.1, γ_V and β_V are monotone functions of V (increasing and decreasing, respectively), and we write

$$\gamma_V \uparrow \gamma \text{ and } \beta_V \downarrow \beta \text{ as } V \downarrow c.$$

In the next theorem, we shall see that upper and lower semicontinuity relate to \limsup and \liminf as continuity relates to limit.

5.5.16. *Lemma.* If $f : X \rightarrow \overline{\mathbb{R}}$ and $c \in X$, then

$$\liminf_{x \rightarrow c} f(x) \leq f(c) \leq \limsup_{x \rightarrow c} f(x).$$

Proof. For any two neighborhoods V and W of c , we have

$$\inf_{x \in V} f(x) \leq f(c) \leq \sup_{x \in W} f(x).$$

In the notations of Definition 5.5.15, $f(c)$ is an upper bound for the γ_V , therefore $\gamma \leq f(c)$; similarly $f(c)$ is a lower bound for the β_W , so $f(c) \leq \beta$. \diamond

5.5.17. Theorem. If $f : X \rightarrow \overline{\mathbb{R}}$ and $c \in X$, the following conditions are equivalent:

- (a) f is lower semicontinuous at c ;
- (b) $\liminf_{x \rightarrow c} f(x) = f(c)$.

Proof. (a) \Rightarrow (b): Let

$$m = \liminf_{x \rightarrow c} f(x);$$

in view of the lemma, we need only show that $m \geq f(c)$. This is trivial if $m = +\infty$ or if $f(c) = -\infty$, so we can suppose that $m < +\infty$ and $f(c) > -\infty$.

If $f(c) > r \in \mathbb{R}$ then, by (a), the set $V = \{x : f(x) > r\}$ is a neighborhood of c , so that

$$r \leq \inf_{x \in V} f(x) \leq m$$

(by the definition of m as a sup of infs); thus $r \leq m$ for every real number $r < f(c)$, therefore $f(c) \leq m$.

(b) \Rightarrow (a): If $f(c) = -\infty$ then (a) holds trivially. Suppose $f(c) > -\infty$. Assuming $f(c) > r \in \mathbb{R}$, we have to show that the set $W = \{x : f(x) > r\}$ is a neighborhood of c . Citing (b), we have

$$\sup_V \left(\inf_{x \in V} f(x) \right) > r,$$

where V runs over the set of all neighborhoods of c , thus there exists a neighborhood V such that

$$\inf_{x \in V} f(x) > r;$$

then $V \subset \{x : f(x) > r\} = W$, therefore W is also a neighborhood of c . \diamond

Dually,

5.5.18. Corollary. If $f : X \rightarrow \overline{\mathbb{R}}$ and $c \in X$, the following conditions are equivalent:

- (a) f is upper semicontinuous at c ;
- (b) $\limsup_{x \rightarrow c} f(x) = f(c)$.

5.5.19. Corollary. If $f : X \rightarrow \overline{\mathbb{R}}$ and $c \in X$, the following conditions are equivalent:

- (a) f is continuous at c ;
- (b) $\liminf_{x \rightarrow c} f(x) = \limsup_{x \rightarrow c} f(x)$.

When the conditions are verified, the number in (b) is equal to $f(c)$.

Proof. (a) \Rightarrow (b): Immediate from 5.5.9, 5.5.17 and the preceding corollary.

(b) \Rightarrow (a): In view of 5.5.16, it is immediate from (b) that

$$\liminf_{x \rightarrow c} f(x) = f(c) = \limsup_{x \rightarrow c} f(x);$$

thus, f is both lower and upper semicontinuous at c (5.5.17 and 5.5.18), hence continuous at c (5.5.9). \diamond

The applications of the next theorem to derivatives will play an important role in the proof that indefinite integrals are a.e. antiderivatives (§5.9). Recall that if $f, g : X \rightarrow \overline{\mathbb{R}}$ then $f + g$ is defined except at the points x where $f(x)$ and $g(x)$ are both infinite and of opposite signs ((1.15.4, (iv))).

5.5.20. Theorem. Let $f, g : X \rightarrow \overline{\mathbb{R}}$ be functions such that $f + g$ is everywhere defined on X and let c be any point of X . Then:

$$(i) \liminf_{x \rightarrow c} (f + g)(x) \geq \liminf_{x \rightarrow c} f(x) + \liminf_{x \rightarrow c} g(x),$$

$$(ii) \limsup_{x \rightarrow c} (f + g)(x) \leq \limsup_{x \rightarrow c} f(x) + \limsup_{x \rightarrow c} g(x),$$

provided that the right members are defined.

Proof. The stipulation at the end of the statement is that the (undefined) sums $(+\infty) + (-\infty)$ and $(-\infty) + (+\infty)$ do not occur on the right side.

It will suffice to prove (i), for (ii) can then be deduced by applying (i) to $-f$ and $-g$. Define

$$\alpha = \liminf_{x \rightarrow c} f(x), \quad \beta = \liminf_{x \rightarrow c} g(x), \quad \gamma = \liminf_{x \rightarrow c} (f + g)(x).$$

By assumption $\alpha + \beta$ is defined; the problem is to show that

$$(*) \quad \gamma \geq \alpha + \beta.$$

Let us first dispose of some special cases: the inequality $(*)$ is trivial if $\gamma = +\infty$, or if one (or both) of α, β is $-\infty$. Thus we can suppose that

$$\alpha > -\infty, \quad \beta > -\infty, \quad \gamma < +\infty.$$

For every neighborhood V of c , write

$$\alpha_V = \inf_{x \in V} f(x), \quad \beta_V = \inf_{x \in V} g(x), \quad \gamma_V = \inf_{x \in V} [f(x) + g(x)]$$

(not to be confused with the notations in Definition 5.5.15); thus,

$$\alpha_V \uparrow \alpha, \quad \beta_V \uparrow \beta \quad \text{and} \quad \gamma_V \uparrow \gamma \quad \text{as} \quad V \downarrow c.$$

Since $\alpha > -\infty$ and $\beta > -\infty$, there exist neighborhoods V of c such that $\alpha_V > -\infty$ and $\beta_V > -\infty$; for such V , the sum $\alpha_V + \beta_V$ is defined and

$$(1) \quad +\infty > \gamma \geq \gamma_V \geq \alpha_V + \beta_V > -\infty.$$

Let (V_n) be a sequence of neighborhoods of c such that

$$\alpha_{V_n} > -\infty \quad \text{and} \quad \sup_n \alpha_{V_n} = \alpha;$$

replacing V_n by $V_1 \cap \dots \cap V_n$, we can suppose that

$$V_n \downarrow \quad \text{and} \quad \alpha_{V_n} \uparrow \alpha.$$

Similarly, there exists a sequence (W_n) of neighborhoods of c such that

$$W_n \downarrow \quad \text{and} \quad \beta_{W_n} \uparrow \beta.$$

Replacing both V_n and W_n by $V_n \cap W_n$, we can suppose that

$$\alpha_{V_n} \uparrow \alpha \quad \text{and} \quad \beta_{V_n} \uparrow \beta.$$

By (1), we have

$$(2) \quad +\infty > \gamma \geq \gamma_{V_n} \geq \alpha_{V_n} + \beta_{V_n}$$

for all n . Since the right member of (2) is increasing and bounded above, it is clear that neither α nor β can be $+\infty$, thus both are in \mathbb{R} and passage to the limit in (2) yields $\gamma \geq \alpha + \beta$. \diamond

5.5.21. Corollary. *With notations as in the theorem, suppose that $f + g$ is everywhere defined on X .*

(1) *If f and g are l.s.c. at c , then so is $f + g$.*

(2) *If f and g are u.s.c. at c , then so is $f + g$.*

Proof. (1) In particular, $f + g$ is defined at c . In view of 5.5.17, this means that the right side of (i) in the preceding theorem is defined, and

$$\liminf_{x \rightarrow c} (f + g)(x) \geq f(c) + g(c);$$

the reverse inequality holds by 5.5.16, thus $f + g$ is l.s.c. at c by 5.5.17.

(2) Apply (1) to $-f$ and $-g$. \diamond

We now apply some of this machinery to difference-quotient functions:

5.5.22. Definition. Let $f : [a, b] \rightarrow \mathbb{R}$, $a < b$. For every $x \in [a, b]$, define

$$(\underline{D}f)(x) = \liminf_{t \rightarrow x, t \neq x} \frac{f(t) - f(x)}{t - x},$$

$$(\overline{D}f)(x) = \limsup_{t \rightarrow x, t \neq x} \frac{f(t) - f(x)}{t - x},$$

called, respectively, the **lower derivate** and the **upper derivate** of f at x .

The liminf and limsup in the above definition are applied to a function of t defined on the subset $[a, b] - \{x\}$ of the metric space $[a, b]$, and x is not an isolated point, that is, x is adherent to $[a, b] - \{x\}$; we are thus in

the general framework of 5.3.1. The results of §5.3 are therefore applicable here; in particular:

5.5.23. Theorem. *With notations as in the preceding definition (5.5.22), f is differentiable at $x \in (a, b)$ if and only if*

$$(\underline{D}f)(x) = (\overline{D}f)(x) \in \mathbb{R},$$

in which case its derivative $f'(x)$ is the common value of the upper and lower derivatives of f at x .

Proof. This is immediate from 5.3.4 and the definitions. \diamond

5.5.24. Theorem. *With notations as in 5.5.22 and 5.3.7,*

$$(i) \quad (\underline{D}f)(x) = \min\{(\underline{D}_-f)(x), (\underline{D}_+f)(x)\},$$

$$(ii) \quad (\overline{D}f)(x) = \max\{(\overline{D}_-f)(x), (\overline{D}_+f)(x)\},$$

for every $x \in (a, b)$. At the endpoints a and b ,

$$(iii) \quad (\underline{D}f)(a) = (\underline{D}_+f)(a), \quad (\underline{D}f)(b) = (\underline{D}_-f)(b),$$

$$(iv) \quad (\overline{D}f)(a) = (\overline{D}_+f)(a), \quad (\overline{D}f)(b) = (\overline{D}_-f)(b).$$

Proof. (i) Let $x \in (a, b)$ and write

$$g(t) = \frac{f(t) - f(x)}{t - x} \quad \text{for all } t \in [a, b] - \{x\}.$$

By 5.3.3,

$$(\underline{D}f)(x) = \min\{\alpha \in \overline{\mathbb{R}} : g(t_n) \rightarrow \alpha \text{ with } t_n \neq x, t_n \rightarrow x\},$$

$$(\underline{D}_-f)(x) = \min\{\alpha \in \overline{\mathbb{R}} : g(t_n) \rightarrow \alpha \text{ with } t_n < x, t_n \rightarrow x\},$$

$$(\underline{D}_+f)(x) = \min\{\alpha \in \overline{\mathbb{R}} : g(t_n) \rightarrow \alpha \text{ with } t_n > x, t_n \rightarrow x\};$$

the equality in (i) is immediate from the fact that if $t_n \neq x$ for all n , then either $t_n < x$ frequently or $t_n > x$ frequently (or both).

(ii) The proof is similar to (i).

(iii), (iv) These equalities are immediate from the definitions, since approach is possible only from one side. \diamond

5.5.25. Theorem. *If $f, g : [a, b] \rightarrow \mathbb{R}$, then*

$$\underline{D}(f - g) \geq \underline{D}f - \overline{D}g$$

at every point of $[a, b]$ for which the difference on the right side is defined.

Proof. Write $h = -g$, so that $f - g = f + h$, and let $x \in [a, b]$. For all $t \neq x$,

$$\frac{(f - g)(t) - (f - g)(x)}{t - x} = \frac{f(t) - f(x)}{t - x} + \frac{h(t) - h(x)}{t - x};$$

by the proof of (i) of 5.5.20 (with neighborhoods V replaced by deleted neighborhoods $V - \{x\}$, possible because x is not an isolated point), we have

$$\begin{aligned} & \liminf_{t \rightarrow x, t \neq x} \frac{(f-g)(t) - (f-g)(x)}{t-x} \\ & \geq \liminf_{t \rightarrow x, t \neq x} \frac{f(t) - f(x)}{t-x} + \liminf_{t \rightarrow x, t \neq x} \frac{h(t) - h(x)}{t-x} \end{aligned}$$

at every x for which the right side is defined, thus

$$\underline{D}(f-g) \geq \underline{D}f + \underline{D}(-g) = \underline{D}f - \overline{D}g$$

at every such x . \diamond

5.5.26. Theorem. If $f, g : [a, b] \rightarrow \mathbb{R}$ then

$$D^+(f-g) \geq D^+f - D^+g$$

at every point of $[a, b]$ where D^+g is finite. The same inequality holds with D^+ replaced by D^- and $[a, b]$ by $(a, b]$.

Proof. Suppose $(D^+g)(x) \in \mathbb{R}$. Writing $f = (f-g) + g$, we have

$$\begin{aligned} (D^+f)(x) &= \limsup_{t \rightarrow x+} \frac{f(t) - f(x)}{t-x} \\ &= \limsup_{t \rightarrow x+} \left\{ \frac{(f-g)(t) - (f-g)(x)}{t-x} + \frac{g(t) - g(x)}{t-x} \right\} \\ &\leq \limsup_{t \rightarrow x+} \frac{(f-g)(t) - (f-g)(x)}{t-x} + \limsup_{t \rightarrow x+} \frac{g(t) - g(x)}{t-x} \end{aligned}$$

by the proof of (ii) of 5.5.20 (with neighborhoods replaced by deleted right neighborhoods), valid because the sum on the right side of the inequality is obviously defined. Thus

$$(D^+f)(x) \leq [D^+(f-g)](x) + (D^+g)(x);$$

since the last term on the right side is finite, it can be transposed to yield the desired inequality. The second assertion of the theorem follows on replacing deleted right neighborhoods by deleted left neighborhoods in the foregoing argument. \diamond

Exercise

1. With notations as in Theorem 5.5.23, f is right-differentiable at a if and only if $(\underline{D}f)(a)$, $(\overline{D}f)(a)$ are equal and finite, in which case $f'_r(a)$ is the common value of the upper and lower derivates of f at a . Similarly for left-differentiability at b .

5.6. Semicontinuous Approximations of Integrable Functions

Throughout this section, λ denotes Lebesgue measure either on \mathbb{R} or on the closed interval $[a, b]$; $\mathcal{L}^1 = \mathcal{L}^1([a, b], \lambda)$ is the class of Lebesgue-integrable functions $f : [a, b] \rightarrow \mathbb{R}$. If $A \subset [a, b]$ we write φ_A for the characteristic function of A , as a function on $[a, b]$.

The following approximation theorem is for application in §5.9 (in the proof that the indefinite integral of $f \in \mathcal{L}^1$ has derivative $f(x)$ almost everywhere):

5.6.1. Theorem. *If $f \in \mathcal{L}^1$ and $\epsilon > 0$, there exist functions $h \in \mathcal{L}^1$ and $k : [a, b] \rightarrow \mathbb{R} \cup \{+\infty\}$ with the following properties:*

- (i) $f \leq h$ a.e.,
- (ii) $\int h d\lambda \leq \int f d\lambda + \epsilon$,
- (iii) k is lower semicontinuous and $f \leq k$ everywhere on $[a, b]$,
- (iv) $h = k$ a.e.

Proof. Informally, every integrable function f admits a lower semicontinuous ‘cover’ k that is equal a.e. to an integrable function h whose integral is as close as we like to that of f .

The proof is by reduction to special cases. We consider, successively, (a) $f = \varphi_E$ the characteristic function of a measurable set $E \subset [a, b]$; (b) f simple and ≥ 0 ; (c) f integrable and ≥ 0 ; and (d) f integrable (the general case).

(a) Suppose $f = \varphi_E$, E a Lebesgue-measurable subset of $[a, b]$. By the regularity of Lebesgue measure on \mathbb{R} (2.4.14, 2.4.18) and the finiteness of $\lambda(E)$, there exist a closed set K and an open set U in \mathbb{R} such that

$$K \subset E \subset U \quad \text{and} \quad \lambda(U - K) < \epsilon$$

(of course U is not required to be a subset of $[a, b]$). As noted in the preceding section, the characteristic function of U (as a function on \mathbb{R}) is l.s.c. on \mathbb{R} , therefore its restriction to $[a, b]$ is l.s.c. on the (metric) topological space $[a, b]$; we denote the restricted function by h , thus $h = \varphi_{U \cap [a, b]}$. Writing $A = U \cap [a, b]$, we have $K \subset E \subset A$, therefore

$$\varphi_K \leq \varphi_E = f \leq h = \varphi_A$$

and

$$\begin{aligned} \int h d\lambda &= \lambda(A) \leq \lambda(U) \leq \lambda(K) + \lambda(U - K) \\ &< \lambda(K) + \epsilon \leq \lambda(E) + \epsilon = \int f d\lambda + \epsilon; \end{aligned}$$

setting $k = h$, the requirements of the theorem are fulfilled. Note also that φ_K is u.s.c., $\varphi_K \leq f$ and

$$\int \varphi_K d\lambda = \lambda(K) \geq \lambda(U) - \lambda(U - K) > \lambda(U) - \epsilon \geq \int h d\lambda - \epsilon \geq \int f d\lambda - \epsilon,$$

thus

$$\int \varphi_K d\lambda > \int f d\lambda - \epsilon.$$

(b) Suppose f is simple and ≥ 0 . Write

$$f = c_1 \varphi_{E_1} + \dots + c_n \varphi_{E_n},$$

where the E_i are pairwise disjoint Lebesgue-measurable subsets of $[a, b]$ and $c_i > 0$ for all i . For each i , choose a closed set K_i and an open set U_i in \mathbb{R} such that

$$K_i \subset E_i \subset U_i \quad \text{and} \quad \lambda(U_i - K_i) < \epsilon / nc_i.$$

Write $A_i = U_i \cap [a, b]$ and let

$$g = c_1 \varphi_{K_1} + \dots + c_n \varphi_{K_n}, \quad h = c_1 \varphi_{A_1} + \dots + c_n \varphi_{A_n}.$$

Then $0 \leq g \leq f \leq h$, h is l.s.c. (5.5.21) and

$$\int (h - g) d\lambda = \sum_{i=1}^n c_i \lambda(A_i - K_i) \leq \sum_{i=1}^n c_i \lambda(U_i - K_i) < \sum_{i=1}^n \epsilon / n = \epsilon;$$

setting $k = h$, the requirements of the theorem are met. Note also that g is u.s.c. and

$$\int g d\lambda > \int f d\lambda - \epsilon.$$

(c) Suppose $f \geq 0$ (and $f \in \mathcal{L}^1$). Choose a sequence (f_n) of (integrable) simple functions such that $0 \leq f_n \uparrow f$. By the preceding case, there exist simple functions g_n and h_n such that

$$0 \leq g_n \leq f_n \leq h_n,$$

g_n is u.s.c., h_n is l.s.c. and

$$\int (h_n - g_n) d\lambda < \epsilon / 2^n.$$

Define

$$G_n = \sup(g_1, \dots, g_n), \quad H_n = \sup(h_1, \dots, h_n);$$

G_n is u.s.c. and H_n is l.s.c. (5.5.8 and 5.5.4), both are simple functions, and

$$0 \leq g_n \leq G_n \leq \sup(f_1, \dots, f_n) = f_n \leq h_n \leq H_n.$$

In particular, $0 \leq f_n - G_n \leq h_n - g_n$, therefore

$$(1) \quad 0 \leq \int f_n d\lambda - \int G_n d\lambda \leq \int (h_n - g_n) d\lambda < \epsilon / 2^n;$$

it follows that

$$\lim_{n \rightarrow \infty} \int G_n d\lambda = \lim_{n \rightarrow \infty} \int f_n d\lambda = \int f d\lambda.$$

{The second limit exists by the monotone convergence theorem, so the first limit exists and is equal to it by (1).} It is elementary that

$$0 \leq H_n - G_n \leq \sum_{i=1}^n (h_i - g_i).$$

{The crux of the matter is that if $\alpha_i, \beta_i \in \mathbb{R}$ and $\alpha_i \leq \beta_i$ ($i = 1, \dots, n$), then

$$\max \beta_i - \max \alpha_i \leq \sum_{i=1}^n (\beta_i - \alpha_i);$$

for, if $\max \beta_i = \beta_j$ and $\max \alpha_i = \alpha_k$, then $\alpha_k \geq \alpha_j$ and

$$\beta_j - \alpha_k \leq \beta_j - \alpha_j \leq \sum_{i=1}^n (\beta_i - \alpha_i). \}$$

Thus,

$$0 \leq \int (H_n - G_n) d\lambda \leq \sum_{i=1}^n \int (h_i - g_i) d\lambda < \sum_{i=1}^n \epsilon/2^i < \epsilon,$$

therefore,

$$(2) \quad \int H_n d\lambda < \int G_n d\lambda + \epsilon \leq \int f_n d\lambda + \epsilon \leq \int f d\lambda + \epsilon$$

for all n ; since $H_n \uparrow$, by the monotone convergence theorem there exists an $h \in \mathcal{L}^1$ such that

$$(3) \quad H_n \uparrow h \text{ a.e.,}$$

and since $H_n \geq 0$ for all n , we can suppose (by modifying h on a negligible set, if necessary) that $h \geq 0$ everywhere on $[a, b]$. From (3) we have

$$\int H_n d\lambda \uparrow \int h d\lambda;$$

passing to the limit in (2), we have

$$\int h d\lambda \leq \int f d\lambda + \epsilon.$$

Define $k = \sup H_n$; then $k : [a, b] \rightarrow [0, +\infty]$, k is l.s.c. (5.5.4) and $k = h$ a.e. by (3). Moreover,

$$f = \sup f_n \leq \sup H_n = k,$$

thus $f \leq k$ (everywhere on $[a, b]$). Since $k = h$ a.e., it follows that $f \leq h$ a.e. This completes the proof for the case that $f \geq 0$.

(d) Consider now the general case that $f \in \mathcal{L}^1$. Write $f = f_1 - f_2$ with f_1, f_2 integrable and ≥ 0 . Applying the preceding case (c) to f_1 and $\epsilon/2$, there exist functions $h_1 \in \mathcal{L}^1$, $k_1 : [a, b] \rightarrow [0, +\infty]$, such that $h_1 \geq 0$, k_1 is l.s.c., $f_1 \leq k_1$ everywhere on $[a, b]$, $h_1 = k_1$ a.e. (hence $f_1 \leq h_1$ a.e.) and

$$(4) \quad \int h_1 d\lambda \leq \int f_1 d\lambda + \epsilon/2.$$

Also, applying case (c) to f_2 and $\epsilon/2$, the proof of (c) shows that there exists a simple function g_2 , with $0 \leq g_2 \leq f_2$ and g_2 u.s.c., such that

$$(5) \quad \int (f_2 - g_2) d\lambda < \epsilon/2.$$

Then $-g_2$ is l.s.c., hence so is $k_1 - g_2$ (5.5.21), and

$$f = f_1 - f_2 \leq f_1 - g_2 \leq k_1 - g_2.$$

Define

$$k = k_1 - g_2, \quad h = h_1 - g_2.$$

Then $k : [a, b] \rightarrow \mathbb{R} \cup \{+\infty\}$ is l.s.c., $f \leq k$ (everywhere), $h \in \mathcal{L}^1$, $k = h$ a.e. (because $k_1 = h_1$ a.e.), hence $f \leq h$ a.e.; moreover,

$$h - f = (h_1 - g_2) - (f_1 - f_2) = (h_1 - f_1) + (f_2 - g_2),$$

therefore

$$\int (h - f) d\lambda = \int (h_1 - f_1) d\lambda + \int (f_2 - g_2) d\lambda < \epsilon/2 + \epsilon/2$$

by (4) and (5), thus h and k meet the requirements of the theorem. \diamond

5.6.2. Remark. For each positive integer n , let $\epsilon = 1/n$ and choose functions h_n, k_n satisfying the conditions (i)–(iv) of the theorem. From (i) and (ii) we see that $\int |h_n - f| d\lambda \rightarrow 0$.

A neater way of packaging this result is as follows. Call a function $k : [a, b] \rightarrow \overline{\mathbb{R}}$ *integrable* if there exists a function $h \in \mathcal{L}^1$ such that $k = h$ a.e., and define the *integral* of k to be the integral of h . The theorem can then be stated succinctly as follows: Every $f \in \mathcal{L}^1$ is the limit in mean of a sequence of lower semicontinuous integrable functions that are $\geq f$.

5.7. F. Riesz's "Rising Sun Lemma"

Riesz's lemma (which we shall use once and only once, in the next section) is part of the technical preparation for the proof that indefinite integrals

are a.e. primitives (§5.9). The following structure theorem for open sets in \mathbb{R} is needed before we can state Riesz's lemma:

5.7.1. Lemma. *Every nonempty open set U in \mathbb{R} is the union $U = \bigcup I_n$ of a countable family of pairwise disjoint intervals that are open sets.*

Proof. {The intervals I_n are permitted to be unbounded (possible for at most two values of n); for example, if $U = \mathbb{R} - \{1, 2\}$ then $U = (-\infty, 1) \cup (1, 2) \cup (2, +\infty)$ is the representation of U promised in the lemma. We reserve the term 'open interval' for intervals of type (a, b) with endpoints $a, b \in \mathbb{R}$, whence the locution "intervals that are open sets". The intervals making up such a decomposition of U are unique (Exercise 1), but this fact is not needed in our application.}

For $x, y \in U$, write $x \sim y$ if the closed interval with endpoints x and y is contained in U (equivalently, there exists an interval I such that $x, y \in I \subset U$). The relation \sim is an equivalence relation in U (for transitivity, note that the union of two intervals with a common point is an interval). Let \mathcal{K} be the set of all equivalence classes for \sim . At any rate, the sets in \mathcal{K} are pairwise disjoint.

claim 1: Every $K \in \mathcal{K}$ is an interval.

Given $x, y \in K$ with $x \leq y$, it suffices to show that $[x, y] \subset K$.¹ Since $x \sim y$, we know that $[x, y] \subset U$. If $z \in [x, y]$ then $[x, z] \subset [x, y] \subset U$ shows that $z \sim x \in K$, therefore $z \in K$; thus $[x, y] \subset K$.

claim 2: Every $K \in \mathcal{K}$ is an open set.

Since K is an interval in \mathbb{R} , we need only show that it has no largest element and no smallest element. Assume to the contrary, for example, that K has a largest element b . Since $b \in U$, there exists a $\delta > 0$ such that $[b - \delta, b + \delta] \subset U$; then $b + \delta \sim b \in K$, therefore $b + \delta \in K$, which contradicts the maximality of b .

For each $K \in \mathcal{K}$ choose a rational number $r_K \in K$. Since the sets in \mathcal{K} are pairwise disjoint, $K \mapsto r_K$ is an injective mapping $\mathcal{K} \rightarrow \mathbb{Q}$, whence the countability of \mathcal{K} . {A slightly more formal argument: For each $K \in \mathcal{K}$, $K \cap \mathbb{Q} \neq \emptyset$. Consider the family $(K \cap \mathbb{Q})_{K \in \mathcal{K}}$ of nonempty subsets of \mathbb{Q} ; by the Axiom of Choice, there exists a mapping $f : \mathcal{K} \rightarrow \mathbb{Q}$ such that $f(K) \in K \cap \mathbb{Q}$ for all $K \in \mathcal{K}$. Since the sets $K \cap \mathbb{Q}$ are pairwise disjoint, f is injective, therefore $\text{card } \mathcal{K} \leq \text{card } \mathbb{Q} = \aleph_0$.} \diamond

In the context of a function $g : [a, b] \rightarrow \mathbb{R}$, let us say that a point $x \in (a, b)$ is a *peak point* if the restriction of g to $[x, b]$ takes its maximum value at the left endpoint x , in other words, $g(t) \leq g(x)$ for all $t \in (x, b]$. If $x \in (a, b)$ is *not* a peak point, let us say that x is *topped to the right*; this means that there exists a point $t \in (x, b]$ such that $g(t) > g(x)$.

¹ *First course*, p. 59, Theorem 4.1.4.

5.7.2. Theorem. (F. Riesz's "Rising sun lemma") Let $g : [a, b] \rightarrow \mathbb{R}$ be a continuous function and let E be the set of all points in (a, b) that are "topped to the right",

$$\begin{aligned} E &= \{x \in (a, b) : g(t) > g(x) \text{ for some } t > x\} \\ &= \{x \in (a, b) : \frac{g(t) - g(x)}{t - x} > 0 \text{ for some } t \in (x, b)\}. \end{aligned}$$

Then:

- (i) E is an open set in \mathbb{R} .
- (ii) $E = \emptyset \Leftrightarrow g$ is a decreasing function.
- (iii) If $E \neq \emptyset$ then, writing $E = \bigcup (a_n, b_n)$ as in the lemma, where the (a_n, b_n) are pairwise disjoint, we have $g(a_n) \leq g(b_n)$ for all n .

Proof. {The second formula for E indicates that it is a gauge of the 'slope' of g . In (iii), $g(a_n) = g(b_n)$ for all except possibly one value of n (Exercise 2).}

(i) Assuming $c \in E$ we have to show that c is interior to E . By assumption, there exists a point $t \in (c, b]$ such that $g(t) > g(c)$.

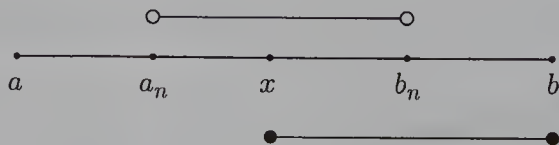


Since g is continuous at c and $g(c) < g(t)$, the values of g remain $< g(t)$ in a neighborhood of c , thus there exists an $\epsilon > 0$ such that $a < c - \epsilon < c + \epsilon < t$ and $g(x) < g(t)$ for all $x \in (c - \epsilon, c + \epsilon)$; then $(c - \epsilon, c + \epsilon) \subset E$ because every $x \in (c - \epsilon, c + \epsilon)$ is topped to the right (at t).

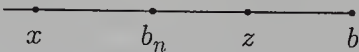
(ii) If g is decreasing, it is obvious that $E = \emptyset$. Assuming g is not decreasing, let us show that $E \neq \emptyset$. By assumption, there exist points x, y with $a \leq x < y \leq b$ and $g(x) < g(y)$; since g is continuous, we can suppose that $a < x < y \leq b$. Then $x \in (a, b)$ is topped to the right (at y), thus $x \in E$.

(iii) Write $E = \bigcup I_n$ as in the lemma. Since $E \subset [a, b]$, the intervals I_n are bounded and $I_n = (a_n, b_n)$ with $a_n, b_n \in [a, b]$. In fact, $a_n, b_n \in [a, b] - E$ (all points of I_n are interior points, so I_n contains neither its own endpoints nor those of the I_m with $m \neq n$).

Fix an index n ; we are to show that $g(a_n) \leq g(b_n)$. Let $x \in (a_n, b_n)$; by the continuity of g , it will suffice to show that $g(x) \leq g(b_n)$ (then let $x \rightarrow a_n +$).



The restriction $g|_{[x,b]}$ has a largest value, say at $z \in [x,b]$. In particular, z is not topped to the right, so $z \notin E$; but $[x,b_n) \subset (a_n,b_n) \subset E$, therefore $z \notin [x,b_n)$, consequently $z \in [b_n,b]$.

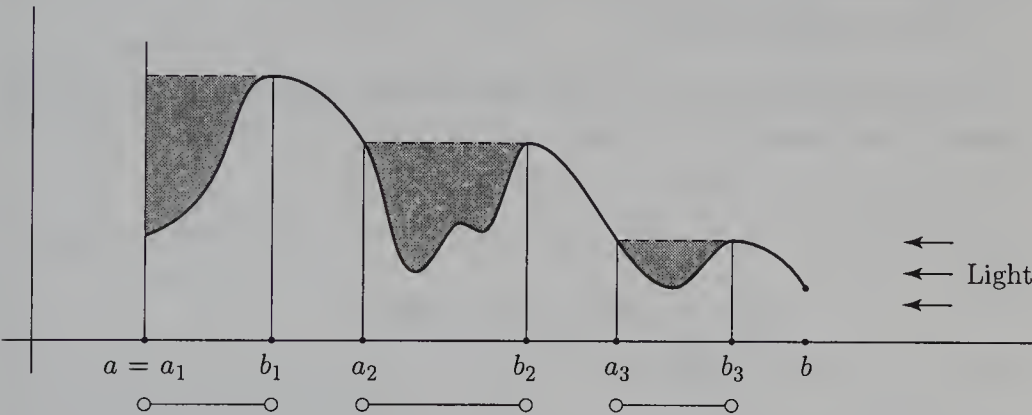


Since $x \in (a_n,b_n)$ and $z \in [b_n,b]$, we have

$$a \leq a_n < x < b_n \leq z \leq b,$$

and, by the choice of z , $g(x) \leq g(z)$ and $g(b_n) \leq g(z)$. Necessarily $g(b_n) = g(z)$. {For, $g(b_n) < g(z)$ would entail $b_n < z$, thus b_n would be topped to the right (at z), contrary to $b_n \notin E$.} Thus $g(x) \leq g(z) = g(b_n)$, so that $g(x) \leq g(b_n)$, as we wished to show. \diamond

Here is a picture illustrating the “Rising sun lemma”², in which the open set E is the union of three open intervals:



The (fanciful) reason for the fanciful name of this result: Think of the graph of the function as the contour of the cross-section, by the xy -plane, of some hills (without ‘overhangs’). Imagine light rays from a distant source (sunrise at $+\infty$?) coming in parallel to the x -axis; the intervals (a_n,b_n) delimit the hollows that remain in shade at the moment of sunrise. As the sun rises higher and light comes down at an angle, parts of the hollows will become illuminated, with nothing remaining in shade when the sun is directly overhead. As long as we’re overhead, let’s imagine a deluge; lakes will form in the hollows (better erect a tall wall at $x = a$ to keep water from leaking off to the left!), spilling over to the right at the points b_n .

²I learned the name from the book of E. Asplund and L. Bungart [*A first course in integration*, Holt, Rinehart and Winston, New York, 1966], p. 268. Cf. B. Sz.-Nagy, *Introduction to real functions and orthogonal expansions* [Oxford, New York, 1965], pp. 107-108.

Exercises

1. If $U = \bigcup I_n = \bigcup J_m$ are two representations of an open set U in the sense of Lemma 5.7.1, then the J 's are, in some order, equal to the I 's.

{Hint: We can suppose that the I_n are the intervals constructed in the proof of the lemma. It suffices to show that each J_m is equal to some I_n . Keeping in mind that endpoints of intervals in \mathbb{R} may be $\pm\infty$, note that no endpoint of J_m can belong to U (such an endpoint would create overlap between two distinct J 's). Next, show that if $J_m \cap I_n \neq \emptyset$ then $J_m \subset I_n$. Finally, show that if the inclusion $J_m \subset I_n$ were proper, then J_m would have an endpoint in I_n .}

2. If $E = \bigcup (a_n, b_n)$ as in the "Rising sun lemma", $g(a_n) < g(b_n)$ implies that $a_n = a$ (hence can happen for at most one index n).

{Hint: If $a_n > a$ then $a_n \in (a, b) - E$, so a_n is a "peak point".}

5.8. Estimates of the Growth of a Continuous Increasing Function

The following theorem (for application in the next section) gives an estimate of the growth of a continuous increasing function in terms of its derivatives:

5.8.1. Theorem. If $f : [a, b] \rightarrow \mathbb{R}$ is continuous and increasing, $r > 0$ and

$$E = \{x \in (a, b) : (D^+ f)(x) > r\},$$

then $f(b) - f(a) \geq r\lambda^*(E)$.

Proof. Here λ^* is Lebesgue outer measure. {With some pains, it can be shown that E is a Borel set (§5.11); the nice thing about outer measure is that it does not care.}

Let G be the set of all $x \in (a, b)$ such that

$$\frac{f(t) - f(x)}{t - x} > r$$

for some $t \in [a, b]$ with $t > x$ (that is, the point on the graph at x is the left end of some chord of slope $> r$). Define $g : [a, b] \rightarrow \mathbb{R}$ by

$$g(x) = f(x) - rx \quad (x \in [a, b]);$$

for $t > x$, we have

$$\frac{f(t) - f(x)}{t - x} > r \Leftrightarrow f(t) - f(x) > r(t - x) \Leftrightarrow g(t) > g(x),$$

therefore

$$G = \{x \in (a, b) : g(t) > g(x) \text{ for some } t \in [a, b] \text{ with } t > x\}.$$

We are now in the framework of the "Rising sun lemma" of the preceding section; in particular, G is an open set in \mathbb{R} .

If $G = \emptyset$ then g is decreasing, therefore $D^+g \leq 0$ on $[a, b]$; then $D^+f \leq r$ on $[a, b]$, so $E = \emptyset$ and the assertion of the theorem is trivially verified.

If $G \neq \emptyset$, then

$$G = \bigcup (a_n, b_n)$$

for a countable (possibly finite) family of pairwise disjoint open intervals such that $g(a_n) \leq g(b_n)$ for all n , that is,

$$r(b_n - a_n) \leq f(b_n) - f(a_n) \quad \text{for all } n.$$

It follows that for each n ,

$$(*) \quad r \sum_{i=1}^n (b_i - a_i) \leq \sum_{i=1}^n [f(b_i) - f(a_i)] \leq f(b) - f(a).$$

{The second inequality is obtained by inserting into the sum non-negative terms corresponding to the gaps between the intervals $(a_1, b_1), \dots, (a_n, b_n)$, then telescoping the sum. For example, assuming $(a_1, b_1), \dots, (a_n, b_n)$ to be arranged from left to right on the number line, $f(a_2) - f(b_1) \geq 0$, $f(a_3) - f(b_2) \geq 0$, etc., by the monotonicity of f .} Since $(*)$ holds for every n , we have

$$r \sum (b_i - a_i) \leq f(b) - f(a),$$

that is,

$$f(b) - f(a) \geq r\lambda(G).$$

To complete the proof, we need only show that $G \supset E$.

Let $x \in E$. By the definition of E ,

$$\limsup_{t \rightarrow x+} \frac{f(t) - f(x)}{t - x} > r.$$

By Theorem 5.3.3, there exists a sequence (t_n) such that $t_n > x$, $t_n \rightarrow x$ and

$$\frac{f(t_n) - f(x)}{t_n - x} \rightarrow (D^+f)(x) > r;$$

if n is any index for which

$$\frac{f(t_n) - f(x)}{t_n - x} > r,$$

then $g(t_n) > g(x)$, therefore $x \in G$. \diamond

5.8.2. *Remark.* A similar argument shows that if

$$F = \{x \in (a, b) : (D^-f)(x) > r\}$$

then $f(b) - f(a) \geq r\lambda^*(F)$. {One could also deduce this by applying the theorem to the function f^* of the proof of 5.4.3.}

5.8.3. Corollary. If $f : [a, b] \rightarrow \mathbb{R}$ is continuous and increasing, $r > 0$ and

$$A = \{x \in (a, b) : (\overline{D}f)(x) > r\},$$

then $f(b) - f(a) \geq \frac{1}{2}r\lambda^*(A)$.

Proof. {Again, A can be shown to be a Borel set, but this is not needed for our application in the next section.} From 5.5.24, we know that for every $x \in (a, b)$,

$$(\overline{D}f)(x) = \max\{(D^+f)(x), (D^-f)(x)\},$$

thus

$$\begin{aligned} A &= \{x \in (a, b) : (D^+f)(x) > r \text{ or } (D^-f)(x) > r\} \\ &= \{x : (D^+f)(x) > r\} \cup \{x : (D^-f)(x) > r\}; \end{aligned}$$

in the notations of the theorem and the remark following it, $A = E \cup F$, therefore

$$\lambda^*(A) \leq \lambda^*(E) + \lambda^*(F) \leq 2[f(b) - f(a)]/r. \diamond$$

5.9. Indefinite Integrals are a.e. Primitives

A function $F : [a, b] \rightarrow \mathbb{R}$ is called an **a.e. primitive** of $f : [a, b] \rightarrow \mathbb{R}$ if, for almost every $x \in (a, b)$, F is differentiable at x and $F'(x) = f(x)$. The purpose of this section is to prove that if f is Lebesgue-integrable then its indefinite integral F is an a.e. primitive of f .

The context: λ is Lebesgue measure on the closed interval $[a, b]$, $a < b$, and $\mathcal{L}^1 = \mathcal{L}^1([a, b], \lambda)$ is the class of all Lebesgue-integrable functions on $[a, b]$.

5.9.1. Lemma. Suppose $h \in \mathcal{L}^1$ and $k : [a, b] \rightarrow \overline{\mathbb{R}}$ is a lower semicontinuous function such that $k = h$ a.e. Let H be the indefinite integral of h ,

$$H(x) = \int_a^x h d\lambda \quad (a \leq x \leq b).$$

Then $\underline{D}H \geq k$ everywhere on $[a, b]$.

Proof. Recall (Definition 5.5.22) that, for all $x \in (a, b)$,

$$(\underline{D}H)(x) = \liminf_{t \rightarrow x, t \neq x} \frac{H(t) - H(x)}{t - x}.$$

By 5.5.24,

$$(\underline{D}H)(x) = \min\{(D_-H)(x), (D_+H)(x)\}$$

for all $x \in (a, b)$, and

$$(\underline{D}H)(a) = (D_+H)(a), \quad (\underline{D}H)(b) = (D_-H)(b).$$

Given $x \in [a, b]$, we are to show that $(\underline{D}H)(x) \geq k(x)$. We can suppose that $k(x) > -\infty$.

Suppose first that $a \leq x < b$. Let r be a real number such that $k(x) > r$. Since k is l.s.c. on $[a, b]$, we know (Definition 5.5.1) that the set

$$V = \{t \in [a, b] : k(t) > r\}$$

is a neighborhood of x in the (metric) topological space $[a, b]$. For almost all $t \in V \cap (x, +\infty)$ we have $h(t) = k(t) > r$, therefore

$$\int_x^t h \, d\lambda \geq r(t - x)$$

for every $t \in V \cap (x, +\infty)$, that is, $H(t) - H(x) \geq r(t - x)$ for all $t \in V \cap (x, +\infty)$. Thus

$$\frac{H(t) - H(x)}{t - x} \geq r \quad \text{for all } t \in V \cap (x, +\infty);$$

it follows from the definition of \liminf (as a sup of infs) that

$$r \leq \inf_{t \in V \cap (x, +\infty)} \frac{H(t) - H(x)}{t - x} \leq \liminf_{t \rightarrow x+} \frac{H(t) - H(x)}{t - x} = (D_+H)(x).$$

Since this is true for every $r < k(x)$, we conclude that

$$(1) \quad k(x) \leq (D_+H)(x) \quad \text{for all } x \in [a, b).$$

By a similar argument, with $(x, +\infty)$ replaced by $(-\infty, x)$, we see that

$$(2) \quad k(x) \leq (D_-H)(x) \quad \text{for all } x \in (a, b].$$

From (1), (2) and the formulas of 5.5.24 cited at the beginning of the proof, it is clear that

$$k(x) \leq (\underline{D}H)(x)$$

for every $x \in [a, b]$. {The utility of semicontinuous approximation is clear from this proof: an inequality $h(x) > r$ conveys no information about points other than x , but semicontinuity projects the inequality $k(x) > r$ into an entire neighborhood of x , and carries h along with it almost everywhere.} \diamond

5.9.2. *Remark.* Note (for use in §5.12) that the ‘dual’ of the lemma is also true: If $h \in \mathcal{L}^1$, $k : [a, b] \rightarrow \overline{\mathbb{R}}$ is upper semicontinuous and $k = h$ a.e., then the function $H : [a, b] \rightarrow \mathbb{R}$ defined by

$$H(x) = \int_a^x h \, d\lambda$$

satisfies $\overline{D}H \leq k$ everywhere on $[a, b]$. {It suffices to apply the lemma to the functions $-h$ and $-k$ (integrable and l.s.c., respectively), noting that $-H$ is the indefinite integral of $-h$.}

5.9.3. Theorem. If $f: [a, b] \rightarrow \mathbb{R}$ is Lebesgue-integrable and $F: [a, b] \rightarrow \mathbb{R}$ is defined by

$$F(x) = \int_a^x f d\lambda \quad (a \leq x \leq b),$$

then $F' = f$ a.e.

Proof. The claim is that, for almost every $x \in (a, b)$, F is differentiable at x and $F'(x) = f(x)$.

The first target is to prove that $\underline{D}F \geq f$ a.e. We have to show that the set

$$A = \{x : (\underline{D}F)(x) - f(x) < 0\}$$

is negligible; given $r \in \mathbb{R}$, $r > 0$, it will suffice to show that the set

$$A_r = \{x : (\underline{D}F)(x) - f(x) < -r\}$$

is negligible (A is the union of the sequence of sets $A_{1/n}$ for $n = 1, 2, 3, \dots$). Given any $\epsilon > 0$, it is enough to show that $\lambda^*(A_r) < \epsilon$.

With the notations ϵ and r established above, by Theorem 5.6.1 there exist functions $h \in \mathcal{L}^1$ and $k: [a, b] \rightarrow \overline{\mathbb{R}}$, with k lower semicontinuous, such that $h = k$ a.e., $f \leq k$ everywhere on $[a, b]$, and

$$(1) \quad \int_a^b h d\lambda < \int_a^b f d\lambda + \frac{1}{2}r\epsilon.$$

Let $H: [a, b] \rightarrow \mathbb{R}$ be the indefinite integral of h ,

$$H(x) = \int_a^x h d\lambda \quad (a \leq x \leq b).$$

Since $h - f \geq 0$ a.e. and $H - F$ is the indefinite integral of $h - f$, it follows that $H - F$ is an increasing function. Let

$$E_r = \{x \in [a, b] : [\overline{D}(H - F)](x) > r\}$$

(note that E_r is a function of H , therefore of ϵ); since $H - F$ is continuous and increasing,

$$(H - F)(b) - (H - F)(a) \geq \frac{1}{2}r\lambda^*(E_r)$$

by Corollary 5.8.3 of the preceding section, in other words

$$(2) \quad (H - F)(b) \geq \frac{1}{2}r\lambda^*(E_r).$$

On the other hand, the inequality of (1) says that

$$(H - F)(b) < \frac{1}{2}r\epsilon;$$

combined with (2), this shows that

$$(3) \quad \lambda^*(E_r) < \epsilon.$$

Now, E_r is a function of ϵ but A_r is not. We are going to show that $A_r \subset E_r$, whence $\lambda^*(A_r) \leq \lambda^*(E_r) < \epsilon$, completing the proof that A is negligible. Arguing contrapositively, assuming $x \notin E_r$ we will show that $x \notin A_r$, that is,

$$(\underline{D}F)(x) - f(x) \geq -r.$$

Consider $F = H - (H - F)$; by 5.5.25,

$$(4) \quad \underline{D}F \geq \underline{D}H - \overline{D}(H - F)$$

at every point of $[a, b]$ where the difference on the right side is defined. Since $H - F$ is increasing and $x \notin E_r$, we have

$$(5) \quad 0 \leq [\overline{D}(H - F)](x) \leq r,$$

thus the subtraction in (4) is permissible at x and it follows from (4) and (5) that

$$(\underline{D}F)(x) \geq (\underline{D}H)(x) - [\overline{D}(H - F)](x) \geq (\underline{D}H)(x) - r;$$

but $(\underline{D}H)(x) \geq k(x)$ by the lemma, and $k(x) \geq f(x)$ by the choice of k , thus

$$(\underline{D}F)(x) \geq f(x) - r,$$

as we wished to show.

We now know that A is negligible, in other words,

$$(6) \quad \underline{D}F \geq f \text{ a.e.}$$

Applying (6) to $-f$ (whose indefinite integral is $-F$), we have

$$\underline{D}(-F) \geq -f \text{ a.e.,}$$

that is,

$$-(\overline{D}F) \geq -f \text{ a.e.,}$$

thus

$$(7) \quad \overline{D}F \leq f \text{ a.e.}$$

From (6), (7) and 5.3.3, we have

$$f \leq \underline{D}F \leq \overline{D}F \leq f \text{ a.e.,}$$

therefore

$$\underline{D}F = \overline{D}F = f \text{ a.e.;}$$

in view of 5.5.23, this means that for almost every $x \in (a, b)$, F is differentiable at x and $F'(x) = f(x)$. \diamond

5.9.4. Corollary. If $F : [a, b] \rightarrow \mathbb{R}$ is absolutely continuous, then F is differentiable almost everywhere.

Proof. By the representation theorem 5.2.1, the function $F - F(a)1$ is an indefinite integral. \diamond

Remarks. In many expositions, the almost everywhere differentiability of absolutely continuous functions precedes that of indefinite integrals, that is, the Corollary precedes (and is a lemma to) the Theorem. The strategy in such expositions is to prove the Corollary directly, without the aid of the machinery of integration theory. The direct proofs employ instead a recursive application of the ‘Rising sun lemma’ (to prove that every continuous function of bounded variation is differentiable a.e.¹) or Vitali’s covering theorem (to prove that every function of bounded variation is differentiable a.e.²); once one renounces the use of integration techniques, the full force of absolute continuity seems to provide no simplification.

The key to the approach in this chapter is (a) an early proof that an AC function maps negligible sets to negligible sets (5.1.16), (b) use of the Radon-Nikodym theorem to represent AC functions via indefinite integrals (5.2.1), and (c) the use of semicontinuous approximations to simplify the proof of the a.e. differentiability of indefinite integrals³.

5.10. Lebesgue’s “Fundamental Theorem of Calculus”

This section is essentially a “Scholium”, a gathering together of earlier results into a memorable form. As in the preceding section, the context is Lebesgue measure on an interval $[a, b]$.

5.10.1. Theorem. The following conditions on a function $f : [a, b] \rightarrow \mathbb{R}$ are equivalent:

- (a) f is Lebesgue-integrable;
- (b) there exists an absolutely continuous function $F : [a, b] \rightarrow \mathbb{R}$ such that $F' = f$ a.e.

When the conditions are satisfied, necessarily

$$F(x) = F(a) + \int_a^x f \, d\lambda \quad \text{for all } x \in [a, b],$$

¹ B. Sz.-Nagy, *Introduction to real functions and orthogonal expansions* [Oxford, 1965], pp. 107, 204.

² H. L. Royden, *Real analysis* [3rd edn., Macmillan, 1988], p. 106, proof of Lemma 9; also E. Hewitt and K. Stromberg, *Real and abstract analysis* [Springer, 1965], p. 264, 17.12 and p. 275, 18.3. Reversing the order of events in these expositions, a proof of the almost-everywhere differentiability of a function of bounded variation, based on Theorem 5.9.3, is given in §12 of the present chapter (Corollary 5.12.8 below).

³ Following E. J. McShane, *Integration* [Princeton, 1944], p. 198, 33.3.

F is uniquely determined by f up to an additive constant, and

$$\int_a^b f \, d\lambda = F(b) - F(a).$$

Proof. (a) \Rightarrow (b): The indefinite integral F of f meets the requirements of (b) by 5.2.1 and the theorem of the preceding section; if also G is absolutely continuous and $G' = f$ a.e. then $(G - F)' = 0$ a.e., therefore $G - F$ is constant by 5.4.5 (one could also prove this using 5.2.1). Then, for all $x \in [a, b]$,

$$G(x) - F(x) = G(a) - F(a) = G(a),$$

thus

$$G(x) = G(a) + F(x) = G(a) + \int_a^x f \, d\lambda,$$

and in particular

$$G(b) = G(a) + \int_a^b f \, d\lambda.$$

(b) \Rightarrow (a): By 5.2.1, $F - F(a)$ is the indefinite integral of a Lebesgue-integrable function $g : [a, b] \rightarrow \mathbb{R}$, and $F' = g$ a.e. by the theorem of the preceding section; thus $f = F' = g$ a.e., therefore f is also Lebesgue-integrable. \diamond

5.10.2. Theorem. The following conditions on a function $F : [a, b] \rightarrow \mathbb{R}$ are equivalent:

- (a) F is absolutely continuous;
- (b) there exists a Lebesgue-integrable function $f : [a, b] \rightarrow \mathbb{R}$ such that

$$F(x) = F(a) + \int_a^x f \, d\lambda \quad \text{for all } x \in [a, b].$$

When the conditions are satisfied, $F' = f$ a.e. and f is a.e. unique (in the sense that any two such functions f are equal a.e.).

Proof. The equivalence of (a) and (b), and the essential uniqueness of f , are proved in 5.2.1. With notations as in (b), $F' = f$ a.e. by the theorem of the preceding section. \diamond

5.11. Measurability of Derivates of a Monotone Function

This section is technical preparation for the Lebesgue decomposition theorem of the next section. The ‘official’ definition of measurable function given in §4.1 requires the function to be real-valued; although the derivates of a monotone function may be infinite-valued, the following theorem¹ shows that they are ‘measurable’ in an appropriate sense:

¹ E. J. McShane, *Integration* [Princeton, 1944], p. 194, 32.1.

5.11.1. Theorem. Let $f : [a, b] \rightarrow \mathbb{R}$ be an increasing function (so that all four derivatives of f are ≥ 0). For each real number k , the function $k \cap D^+ f$ defined by

$$(k \cap D^+ f)(x) = \min\{k, (D^+ f)(x)\} \quad (a \leq x \leq b)$$

is a bounded Borel function (hence is Lebesgue-integrable). The same is true with D^+ replaced by D_+ , D^- or D_- .

Proof. The convention is that $(D^+ f)(b) = (D_+ f)(b) = 0$ and $(D^- f)(a) = (D_- f)(a) = 0$.

To simplify the notations, we extend f to an increasing function on \mathbb{R} by defining $f(x) = f(a)$ for $x < a$ and $f(x) = f(b)$ for $x > b$. (Note that the 'honest' derivatives of the extended function at a and b agree with the preceding conventions.) It will suffice to prove the assertion of the theorem (except for integrability) for the extended function, then restrict to the closed interval $[a, b]$.

For each real number $\alpha > 0$, define functions

$$\varphi_\alpha : \mathbb{R} \rightarrow \overline{\mathbb{R}}, \quad \psi_\alpha : \mathbb{R} \rightarrow \overline{\mathbb{R}}$$

by the formulas

$$\begin{aligned} \varphi_\alpha(x) &= \sup_{t \in (x, x+\alpha)} \frac{f(t) - f(x)}{t - x} \\ &= \sup_{s \in (0, \alpha)} \frac{f(x+s) - f(x)}{s}, \\ \psi_\alpha(x) &= \sup_{r \in \mathbb{Q} \cap (0, \alpha)} \frac{f(x+r) - f(x)}{r}. \end{aligned}$$

It is obvious that $0 \leq \psi_\alpha \leq \varphi_\alpha \leq +\infty$ and that, for each $x \in \mathbb{R}$, $\varphi_\alpha(x)$ and $\psi_\alpha(x)$ are increasing functions of α , so that

$$\varphi_\alpha(x) \downarrow \quad \text{and} \quad \psi_\alpha(x) \downarrow \quad \text{as} \quad \alpha \downarrow.$$

claim 1: $\psi_\alpha = \varphi_\alpha$ for all $\alpha > 0$.

Fix $\alpha > 0$, $x \in \mathbb{R}$ and let $0 < s < \alpha$. For each rational number r such that $s < r < \alpha$, we have $f(x+r) \geq f(x+s)$, thus

$$\psi_\alpha(x) \geq \frac{f(x+r) - f(x)}{r} \geq \frac{f(x+s) - f(x)}{r};$$

letting $r \rightarrow s+$ in the right-hand member, we have

$$\psi_\alpha(x) \geq \frac{f(x+s) - f(x)}{s}, \quad .$$

and the validity of this inequality for all $s \in (0, \alpha)$ implies that $\psi_\alpha(x) \geq \varphi_\alpha(x)$. The reverse inequality was noted earlier.

In the proofs of the following claims, it is sometimes convenient to use the original formula for φ_α , sometimes that for ψ_α .

claim 2: $\varphi_\alpha \rightarrow D^+f$ pointwise as $\alpha \rightarrow 0+$. More precisely, if $\alpha_n > 0$ and $\alpha_n \rightarrow 0$ as $n \rightarrow \infty$, then

$$\varphi_{\alpha_n}(x) \rightarrow (D^+f)(x) \text{ in } \overline{\mathbb{R}} \text{ for every } x \in \mathbb{R}.$$

Fix $x \in \mathbb{R}$. Write $V_n = (x - \alpha_n, x + \alpha_n)$ ($n = 1, 2, 3, \dots$) and $A = (x, +\infty)$; since $\text{diam } V_n = 2\alpha_n \rightarrow 0$, Lemma 5.3.2 is applicable.

From claims 1 and 2, we see that

$$(*) \quad \psi_\alpha \rightarrow D^+f \text{ pointwise as } \alpha \rightarrow 0+.$$

Moreover, $\psi_\alpha \downarrow D^+f$ pointwise as $\alpha \downarrow 0$, in the following sense: if $\alpha_n > 0$, $\alpha_n \downarrow 0$ and $x \in \mathbb{R}$, then the sequence $\psi_{\alpha_n}(x)$ is decreasing, with infimum $(D^+f)(x)$.

claim 3: If $\alpha > 0$ and $c \in \mathbb{R}$, then the set

$$E = \{x : \psi_\alpha(x) \leq c\}$$

is a Borel set.

For every $r \in \mathbb{Q} \cap (0, \alpha)$, define $g_r : \mathbb{R} \rightarrow \mathbb{R}$ by the formula

$$g_r(x) = \frac{f(x+r) - f(x)}{r};$$

note that g_r is a linear combination of two increasing functions of x . It is easy to see that every increasing function on \mathbb{R} is Borel. {For example, every inverse image $f^{-1}((t, +\infty))$ ($t \in \mathbb{R}$) is an interval (because $f(x) > t \Leftrightarrow f(y) > t$ for all $y \geq x$), hence is a Borel set; thus f is Borel.} It follows that g_r is a Borel function. By definition, ψ_α is the upper envelope of the family (g_r) , that is,

$$\psi_\alpha(x) = \sup_{r \in \mathbb{Q} \cap (0, \alpha)} g_r(x) \quad \text{for all } x,$$

therefore

$$\psi_\alpha(x) \leq c \Leftrightarrow g_r(x) \leq c \text{ for all } r \in \mathbb{Q} \cap (0, \alpha);$$

in other words,

$$E = \bigcap_{r \in \mathbb{Q} \cap (0, \alpha)} \{x : g_r(x) \leq c\},$$

thus E is the intersection of a countable family of Borel sets, hence is a Borel set.

It follows from claim 3 that the set

$$\{x : \psi_\alpha(x) < c\} = \bigcup_{n=1}^{\infty} \{x : \psi_\alpha(x) \leq c - 1/n\}$$

is also a Borel set, therefore so is its complement $\{x : \psi_\alpha(x) \geq c\}$.

claim 4: For every real number c , the set

$$\{x : (D^+ f)(x) \geq c\}$$

is a Borel set.

For, by the remarks following claim 2, we have $\psi_{1/n} \downarrow D^+ f$ pointwise as $n \rightarrow \infty$, therefore

$$(D^+ f)(x) \geq c \Leftrightarrow \psi_{1/n}(x) \geq c \text{ for all } n;$$

it follows that

$$\{x : (D^+ f)(x) \geq c\} = \bigcap_{n=1}^{\infty} \{x : \psi_{1/n}(x) \geq c\}$$

is the intersection of a sequence of Borel sets.

Now let $k \in \mathbb{R}$. If $k \leq 0$ then $k \cap D^+ f$ is the constant function k (because $D^+ f \geq 0$), a bounded Borel function in good standing. If $k > 0$ then $0 \leq k \cap D^+ f \leq k$, so $k \cap D^+ f$ is certainly bounded; moreover, for every real number c ,

$$\{x : (k \cap D^+ f)(x) \geq c\} = \{x : k \geq c \text{ and } (D^+ f)(x) \geq c\};$$

this set is empty if $k < c$, and if $k \geq c$ it is equal to $\{x : (D^+ f)(x) \geq c\}$, thus, in view of claim 4, it is always a Borel set.

This completes the proof that if $f : [a, b] \rightarrow \mathbb{R}$ is an increasing function, then $k \cap D^+ f$ is a bounded Borel function for every $k \in \mathbb{R}$. We can infer that $k \cap D^- f$ is Borel by using the ‘*-trick’ of Corollary 5.4.3: writing $x^* = a + b - x$ and $f^*(x) = -f(x^*)$, we have

$$(D^- f)(x^*) = (D^+ f^*)(x) \text{ for all } x \in [a, b].$$

Since f^* is increasing, $k \cap D^+ f^*$ is a Borel function by what we have already proved; since $x \mapsto x^*$ is a homeomorphism of $[a, b]$ onto itself, and

$$\begin{aligned} (k \cap D^- f)(x) &= \min\{k, (D^- f)(x)\} \\ &= \min\{k, (D^+ f^*)(x^*)\} \\ &= (k \cap D^+ f^*)(x^*), \end{aligned}$$

it is clear that $k \cap D^- f$ is also a Borel function.

The analogous assertions for D_+ and D_- are left as exercises (they are not needed for the application in the next section). \diamond

5.11.2. Remark. The conclusion of the theorem is also true for decreasing functions and for continuous (but not necessarily monotone) functions.²

²McShane, loc. cit.

Exercises

1. Every function $f : [a, b] \rightarrow \mathbb{R}$ of bounded variation is a bounded Borel function.
{Hint: Jordan decomposition.}
2. Complete the proof of 5.11.1 for D_+ and D_- .

5.12. The Lebesgue Decomposition of a Function of Bounded Variation

The theorem in question (proved in 5.12.9 below):

Every function $F : [a, b] \rightarrow \mathbb{R}$ of bounded variation can be written as a sum $F = G + H$ with G absolutely continuous and $H' = 0$ a.e. (in particular, F is differentiable a.e.). Such a representation is essentially unique: all others are of the form $F = (G + c1) + (H - c1)$, where $c1$ is a constant function.

Included in this result is Theorem C of the remarks at the beginning of §5.4: Every increasing function $F : [a, b] \rightarrow \mathbb{R}$ is differentiable a.e.

We begin the proof with a general observation on limits. As in Definition 5.3.1, let (X, d) be a metric space, $A \subset B \subset X$, $f : B \rightarrow \mathbb{R}$, $c \in \overline{A}$, and let $g : B \rightarrow \mathbb{R}$ be another function:

X

∪

B

∪

$\xrightarrow{f, g}$

\mathbb{R}

$c \in \overline{A} \supset A$

(In §5.3, the functions were allowed to have infinite values; the motive for requiring finite values is to simplify the algebra.)

5.12.1. Lemma. *With the preceding notations, suppose that g has a finite limit*

$$\lim_{x \rightarrow c, x \in A} g(x) = L \in \mathbb{R}.$$

Then

$$\limsup_{x \rightarrow c, x \in A} [f(x) + g(x)] = \limsup_{x \rightarrow c, x \in A} f(x) + L$$

and similarly with limsup replaced by liminf.

Proof. The assumption $L \in \mathbb{R}$ assures that the sum on the right side exists (and that L can be transposed freely). To simplify the notations, we write briefly

$$\limsup(f + g) = \limsup f + L$$

for the equation to be verified. Let

$$S = \{s \in \overline{\mathbb{R}} : f(x_n) \rightarrow s \text{ for some sequence } x_n \in A \text{ with } x_n \rightarrow c\}.$$

As noted in 5.3.3, $\limsup f$ is the largest element of S . Similarly, $\limsup (f + g)$ is the largest element of the set

$$T = \{t \in \overline{\mathbb{R}} : (f + g)(x_n) \rightarrow t \text{ for some sequence } x_n \in A \text{ with } x_n \rightarrow c\}.$$

If (x_n) is a sequence in A with $x_n \rightarrow c$ and if $t \in \overline{\mathbb{R}}$, then $g(x_n) \rightarrow L$ and

$$f(x_n) + g(x_n) \rightarrow t \Leftrightarrow f(x_n) \rightarrow t - L;$$

thus, $t \in T \Leftrightarrow t - L \in S$, that is,

$$T = \{s + L : s \in S\},$$

and the first assertion of the lemma reduces to the observation that $\max T = \max S + L$. For the second assertion, replace \limsup by \liminf and \max by \min in the preceding argument. \diamond

Here is an application of the preceding lemma to derivatives:

5.12.2. Lemma. If $f, g : [a, b] \rightarrow \mathbb{R}$, $x \in [a, b)$ and g is right-differentiable at x , then

$$\begin{aligned} [D^+(f + g)](x) &= (D^+f)(x) + g'_r(x), \\ [D_+(f + g)](x) &= (D_+f)(x) + g'_r(x). \end{aligned}$$

The analogous relations hold for D^- and D_- , assuming g left-differentiable at a point $x \in (a, b]$.

Proof. For $t \neq x$,

$$\frac{(f + g)(t) - (f + g)(x)}{t - x} = \frac{f(t) - f(x)}{t - x} + \frac{g(t) - g(x)}{t - x};$$

by hypothesis, the second term on the right has a finite limit $g'_r(x)$ as $t \rightarrow x+$, thus the asserted formulas follow from the preceding lemma (with $A = (x, b]$). \diamond

In what follows, λ denotes Lebesgue measure on a closed interval $[a, b]$, and $\mathcal{L}^1 = \mathcal{L}^1([a, b], \lambda)$ is the class of Lebesgue-integrable functions on $[a, b]$.

5.12.3. Lemma. If $g \in \mathcal{L}^1$ is bounded, say $|g| \leq M < +\infty$, and if G is the indefinite integral of g , then $|DG| \leq M$ for $D = D^+, D_+, D^-, D_-$.

Proof. If $x, t \in [a, b]$ and $t > x$, then

$$|G(t) - G(x)| = \left| \int_x^t g \, d\lambda \right| \leq \int_x^t |g| \, d\lambda \leq M(t - x),$$

thus

$$\left| \frac{G(t) - G(x)}{t - x} \right| \leq M,$$

whence it is obvious that $|(D^+G)(x)| \leq M$ and $|(D_+G)(x)| \leq M$. Similarly for the left derivatives at $x \in (a, b]$. \diamond

5.12.4. Lemma. If $F : [a, b] \rightarrow \mathbb{R}$, $g \in \mathcal{L}^1$ is bounded and G is the indefinite integral of g , then

$$D^+(F - G) \geq D^+F - D^+G$$

at every point of $[a, b)$. The same inequality holds with D^+ replaced by D^- and $[a, b)$ by $(a, b]$.

Proof. Since D^+G is finite-valued by the preceding lemma, the subtraction on the right side is permissible at every point of $[a, b)$, thus the asserted inequality holds by Theorem 5.5.26. \diamond

5.12.5. Lemma. If $F : [a, b] \rightarrow \mathbb{R}$ is an increasing function, then there exists a function $f \in \mathcal{L}^1$ such that $D^+F = f$ a.e. (in particular, D^+F is finite a.e.). Moreover, the function $K : [a, b] \rightarrow \mathbb{R}$ defined by

$$K(x) = F(x) - \int_a^x f \, d\lambda \quad (a \leq x \leq b)$$

is increasing and $D^+K = 0$ a.e.

Proof. Since F is increasing, all of its derivatives are ≥ 0 . For every positive integer n , define

$$f_n = n \cap D^+F$$

(with the convention that $(D^+F)(b) = 0$); then $0 \leq f_n \leq n$ and $f_n \in \mathcal{L}^1$ by the theorem of the preceding section.

For each n , let g_n be an upper semicontinuous simple function such that $0 \leq g_n \leq f_n$ and

$$\int_a^b f_n \, d\lambda < \int_a^b g_n \, d\lambda + 1/n$$

(5.6.1, part (c) of the proof). {First choose a simple function s_n with $0 \leq s_n \leq f_n$ and

$$\int (f_n - s_n) d\lambda < 1/2n,$$

then choose an upper semicontinuous simple function g_n with $0 \leq g_n \leq s_n$ and

$$\int (s_n - g_n) d\lambda < 1/2n.$$

Let G_n be the indefinite integral of g_n ,

$$G_n(x) = \int_a^x g_n d\lambda \quad (a \leq x \leq b).$$

claim 1: $F - G_n$ is an increasing function.

This will be proved by verifying the conditions needed to apply (1) of Corollary 5.4.7. Since F is increasing and G_n is continuous (even AC), the function $F - G_n$ has finite one-sided limits at every point of $[a, b]$ and, for every $x \in [a, b]$,

$$\begin{aligned} \lim_{t \rightarrow x+} [F(t) - G_n(t)] &= \lim_{t \rightarrow x+} F(t) - \lim_{t \rightarrow x+} G_n(t) \\ &= \lim_{t \rightarrow x+} F(t) - G_n(x) \\ &\geq F(x) - G_n(x); \end{aligned}$$

interpreting “lim” as “liminf” (permissible by 5.3.4), we see that $F - G_n$ satisfies condition (i) of 5.4.2.

Similarly, for every $x \in (a, b]$,

$$\lim_{t \rightarrow x-} [F(t) - G_n(t)] = \lim_{t \rightarrow x-} F(t) - G_n(x) \leq F(x) - G_n(x),$$

therefore (interpreting “lim” as “limsup”) $F - G_n$ satisfies condition (ii) of 5.4.2.

Finally, for every $x \in (a, b)$,

$$(*) \quad [D^+(F - G_n)](x) \geq (D^+F)(x) - (D^+G_n)(x)$$

by the preceding lemma (recall that g_n is simple, therefore bounded). Since g_n is u.s.c.,

$$(\overline{D}G_n)(x) \leq g_n(x)$$

(Remark 5.9.2), that is,

$$\max\{(D^+G_n)(x), (\overline{D}G_n)(x)\} \leq g_n(x)$$

(Theorem 5.5.24), in particular

$$(**) \quad (D^+G_n)(x) \leq g_n(x).$$

Thus

$$\begin{aligned} [D^+(F - G_n)](x) &\geq (D^+F)(x) - g_n(x) && \text{(by (*) and (**))} \\ &\geq (D^+F)(x) - f_n(x) && \text{(because } g_n \leq f_n \text{)} \\ &\geq 0 && \text{(by the definition of } f_n \text{).} \end{aligned}$$

This shows that the hypothesis in (1) of 5.4.7 is satisfied (with a vengeance, the countable exceptional set being in fact empty) and completes the proof that $F - G_n$ is an increasing function.

claim 2: If $a \leq \alpha < \beta \leq b$ then

$$\int_{\alpha}^{\beta} g_n d\lambda \leq F(\beta) - F(\alpha) \quad \text{for all } n.$$

For, by the preceding claim, $F(\alpha) - G_n(\alpha) \leq F(\beta) - G_n(\beta)$, thus $G_n(\beta) - G_n(\alpha) \leq F(\beta) - F(\alpha)$, which is the asserted inequality (recall that G_n is the indefinite integral of g_n).

For every closed interval $[\alpha, \beta] \subset [a, b]$, it follows from the choice of g_n that

$$0 \leq \int_{\alpha}^{\beta} (f_n - g_n) d\lambda \leq \int_a^b (f_n - g_n) d\lambda < 1/n$$

(the first two inequalities because $f_n - g_n \geq 0$), therefore

$$\int_{\alpha}^{\beta} f_n d\lambda < \int_{\alpha}^{\beta} g_n d\lambda + 1/n \leq F(\beta) - F(\alpha) + 1/n$$

(the last inequality by claim 2); thus

$$(1) \quad \int_{\alpha}^{\beta} f_n d\lambda < F(\beta) - F(\alpha) + 1/n \quad \text{for all } n.$$

Since $f_n \uparrow$ (indeed, $f_n \uparrow D^+F$) it follows from (1) (with $\alpha = a$, $\beta = b$) and the monotone convergence theorem that $f_n \uparrow f$ a.e. for some $f \in \mathcal{L}^1$, therefore $f = D^+F$ a.e. Passing to the limit in (1), we have

$$(2) \quad \int_{\alpha}^{\beta} f d\lambda \leq F(\beta) - F(\alpha) \quad \text{for all } [\alpha, \beta] \subset [a, b].$$

Define a function $K : [a, b] \rightarrow \mathbb{R}$ by the formula

$$(3) \quad K(x) = F(x) - \int_a^x f d\lambda \quad (a \leq x \leq b).$$

{Informally,

$$K(x) = F(x) - \int_a^x (D^+F) d\lambda;$$

in a sense, K is what remains—as we shall see, not much—after ‘exhausting’ F by the indefinite integral of one of its derivatives.}

claim 3: K is an increasing function.

For, if $a \leq \alpha < \beta \leq b$, we see from (3) and (2) that

$$K(\beta) - K(\alpha) = F(\beta) - F(\alpha) - \int_{\alpha}^{\beta} f \, d\lambda \geq 0,$$

thus $K\alpha \leq K(\beta)$.

By Lebesgue’s theorem on primitives (5.9.3), the indefinite integral on the right side of (3) is differentiable at almost every x , with derivative $f(x)$; it follows from Lemma 5.12.2 that, for almost every x ,

$$(D^+K)(x) = (D^+F)(x) - f(x),$$

and, since $D^+F = f$ a.e., we have $D^+K = 0$ a.e. \diamond

5.12.6. *Remark.* Replacing D^+ by D^- in the preceding argument and applying it to any increasing function $K : [a, b] \rightarrow \mathbb{R}$, we see that there exists a function $k \in \mathcal{L}^1$ such that $D^-K = k$ a.e. and such that the function

$$H(x) = K(x) - \int_a^x k \, d\lambda \quad (a \leq x \leq b)$$

is increasing and $D^-H = 0$ a.e.

The following theorem delivers “Theorem C” promised in the preliminary remarks of §5.4:

5.12.7. *Theorem. (Lebesgue decomposition)* If $F : [a, b] \rightarrow \mathbb{R}$ is any increasing function, then

(i) F is differentiable a.e.;

(ii) there exist increasing functions $G, H : [a, b] \rightarrow \mathbb{R}$ such that $F = G + H$, G is absolutely continuous and $H' = 0$ a.e.

Proof. Let K be the increasing function constructed in 5.12.5 and apply 5.12.6 to K : there exists a function $k \in \mathcal{L}^1$ such that the function $H : [a, b] \rightarrow \mathbb{R}$ defined by

$$(4) \quad H(x) = K(x) - \int_a^x k \, d\lambda \quad (a \leq x \leq b)$$

is increasing and $D^-H = 0$ a.e. Combining (4) with the defining formula for K in 5.12.5 (and with f the integrable function constructed there), we have

$$(5) \quad H(x) = F(x) - \int_a^x f \, d\lambda - \int_a^x k \, d\lambda \quad (a \leq x \leq b).$$

Applying the theorem on primitives (5.9.3) to the indefinite integrals on the right side of (5), it follows from Lemma 5.12.2 that

$$D^+H = D^+F - f - k \quad \text{a.e.};$$

since $D^+F = f$ a.e. and $k = D^-K$ a.e., it follows that

$$(6) \quad D^+H = -D^-K \quad \text{a.e.}$$

Since H and K are increasing, both derivatives in (6) are ≥ 0 , so it follows from (6) that

$$D^+H = D^-K = 0 \quad \text{a.e.},$$

therefore $D^+H = k = 0$ a.e. Substituting this data into (5), we have

$$(7) \quad H(x) = F(x) - \int_a^x f \, d\lambda \quad (a \leq x \leq b).$$

Moreover,

$$0 \leq D_+H \leq D^+H = 0 \quad \text{a.e.}$$

(for the second inequality, see 5.3.3), therefore $D_+H = D^+H = 0$ a.e.; thus H is right-differentiable a.e. and $H'_r = 0$ a.e. (5.3.10). On the other hand,

$$0 \leq D_-H \leq D^-H = 0 \quad \text{a.e.}$$

by the construction of H , therefore H is left-differentiable a.e. and $H'_l = 0$ a.e. From the preceding two remarks, we see that H is differentiable a.e. and

$$(8) \quad H' = 0 \quad \text{a.e.};$$

combined with (7) (and the theorem on primitives), this shows that F is differentiable a.e. and $F' = f$ a.e.

If $G : [a, b] \rightarrow \mathbb{R}$ is the function defined by

$$G(x) = \int_a^x f \, d\lambda \quad (a \leq x \leq b)$$

(the indefinite integral of f), formula (7) may be written $F = G + H$, where G is absolutely continuous (5.1.13) and H is increasing with $H' = 0$ a.e.; moreover, $f = F' \geq 0$ a.e., therefore G is also increasing. \diamond

5.12.8. Corollary. *Every function $F : [a, b] \rightarrow \mathbb{R}$ of bounded variation is differentiable a.e.*

Proof. By the Jordan decomposition (5.1.7), F is the difference of two increasing functions. \diamond

5.12.9. Corollary. *Suppose $F : [a, b] \rightarrow \mathbb{R}$ is of bounded variation.*

(1) *One can write $F = G + H$ with G absolutely continuous (hence $H = F - G$ is of bounded variation) and $H' = 0$ a.e.*

(2) If also $F = G_1 + H_1$ with G_1 absolutely continuous and $H_1' = 0$ a.e., then

$$G_1 = G + c1 \quad \text{and} \quad H_1 = H - c1$$

for a suitable constant function $c1$.

Proof. (1) By the Jordan decomposition, we are reduced to the case that F is increasing, i.e., to the theorem.

(2) From $G_1 + H_1 = F = G + H$, we have $G_1 - G = H - H_1$, where $G_1 - G$ is absolutely continuous and $(H - H_1)' = 0$ a.e., therefore $G_1 - G$ is constant (5.4.5). \diamond

Functions that have 0 derivative a.e. and are in some sense 'nice' are called 'singular'. There seems to be no standard terminology. For example, a function $H : [a, b] \rightarrow \mathbb{R}$ with $H' = 0$ a.e. is called 'singular' if H is monotone¹ or if H is continuous². The preceding corollary (5.12.9) suggests the following definition:

5.12.10. *Definition.* A function $H : [a, b] \rightarrow \mathbb{R}$ is said to be **singular** if it is of bounded variation (hence differentiable a.e.) and $H' = 0$ a.e.

5.12.11. *Remark.* With notations as in 5.12.9, G and H can be made *unique* by 'normalizing' G so that $G(a) = 0$ (replace G by $G - G(a)1$ and H by $H + G(a)1$). When G is so normalized, it is the indefinite integral of some Lebesgue-integrable function (5.2.1).

5.12.12. *Definition.* Let $F : [a, b] \rightarrow \mathbb{R}$ be of bounded variation and write $F = G + H$ with G absolutely continuous and H singular (5.12.9). When G is normalized so that $G(a) = 0$ (as in the preceding remark), we call $F = G + H$ the **Lebesgue decomposition** of F ; G is called the **absolutely continuous part** and H the **singular part** of F .

The final corollary is essentially 5.12.9 repackaged in a more memorable form:

5.12.13. *Corollary.* If $F : [a, b] \rightarrow \mathbb{R}$ is any function of bounded variation, there exist unique functions $G, H : [a, b] \rightarrow \mathbb{R}$ such that $F = G + H$, G is an indefinite integral and H is singular. More precisely, G is the indefinite integral of f , where f is a Lebesgue-integrable function (unique a.e.) such that $F' = f$ a.e.

Proof. Write $F = G + H$ as in the preceding definition, and express G as the indefinite integral of a function $f \in \mathcal{L}^1$ (5.2.1). Since H is singular, $F' = f$ a.e. by the theorem on primitives (5.9.3). \diamond

¹ H. L. Royden, *Real analysis* [3rd edn., Macmillan, 1988], p. 111, Exercise 16.

² B. Sz.-Nagy, *Introduction to real functions and orthogonal expansions* [Oxford, 1965], p. 207.

5.12.14. *Remark.* In hindsight, we can make the following observation: If $F : [a, b] \rightarrow \mathbb{R}$ is any function of bounded variation, then F is differentiable a.e., the function $F^\bullet : [a, b] \rightarrow \mathbb{R}$ defined by

$$F^\bullet(x) = \begin{cases} F'(x) & \text{if } F \text{ is differentiable at } x \\ 0 & \text{otherwise} \end{cases}$$

is Lebesgue-integrable (it is equal a.e. to the function f of the preceding corollary), and the function $H : [a, b] \rightarrow \mathbb{R}$ defined by

$$H(x) = F(x) - \int_a^x F^\bullet d\lambda$$

is singular. If, moreover, F is increasing, then so is H (by the proof of 5.12.7).

5.13. Lebesgue's Criterion for Riemann-Integrability

The main result in this section is Lebesgue's characterization of the class of Riemann-integrable functions. Let $f : [a, b] \rightarrow \mathbb{R}$ be a *bounded* function, D its set of discontinuities:

$$D = \{x \in [a, b] : f \text{ is not continuous at } x\}.$$

Lebesgue's theorem: f is Riemann-integrable if and only if D has (Lebesgue) measure zero.¹

The proof we shall give is short and conceptual; some general (and easy) preliminaries on semicontinuity only make it appear long.

Lebesgue's criterion for Riemann-integrability is usually expressed by saying that f is 'continuous a.e.' It does *not* mean that «there exists a continuous function g such that $f = g$ a.e.» (consider the characteristic function f of the subinterval $[a, \frac{1}{2}(a+b)]$); and it does *not* mean «there exists a negligible set N such that the restriction of f to $[a, b] - N$ is continuous» (let $N = \mathbb{Q} \cap [a, b]$ be the set of rational numbers in $[a, b]$ and let f be its characteristic function; the restriction of f to $[a, b] - N$ is identically 0, but f is nowhere continuous). The preceding remarks suggest that Lebesgue's criterion is subtle despite its straightforward outward appearance.

We need only review a bare minimum of the machinery of Riemann integration.² Let $f : [a, b] \rightarrow \mathbb{R}$ be a bounded function, $m = \inf f$, $M = \sup f$. A *subdivision* σ of $[a, b]$ is a finite list of points

$$\sigma = \{a = a_0 < a_1 < \cdots < a_n = b\}$$

¹ For an elementary proof not requiring the Lebesgue integral, see *First course*, p. 211, §11.4.

² For full details, see, e.g., *First course*, Chapter 9.

(beginning at a , strictly increasing, and ending at b). For $i = 1, \dots, n$ we write

$$m_i = \inf\{f(x) : a_{i-1} \leq x \leq a_i\}, \\ M_i = \sup\{f(x) : a_{i-1} \leq x \leq a_i\},$$

called the *bounds* of f over the subinterval $[a_{i-1}, a_i]$. The *lower sum* and *upper sum* of f for the subdivision σ are defined, respectively, by the expressions

$$s(\sigma) = \sum_{i=1}^n m_i(a_i - a_{i-1}), \quad S(\sigma) = \sum_{i=1}^n M_i(a_i - a_{i-1}).$$

The (Darboux) *lower integral* of f over $[a, b]$ is the supremum of the lower sums, denoted

$$\int_a^b f(t)dt = \sup\{s(\sigma) : \sigma \text{ any subdivision of } [a, b]\},$$

and the *upper integral* is the infimum of the upper sums,

$$\int_a^b f(t)dt = \inf\{S(\sigma) : \sigma \text{ any subdivision of } [a, b]\}.$$

If the lower and upper integrals of f are equal, then f is said to be *Riemann-integrable*; their common value is then denoted

$$\int_a^b f(t)dt,$$

called the *Riemann integral* of f over $[a, b]$.

Half of Lebesgue's theorem could have been an exercise in §5.4:

5.13.1. Lemma. *If $f : [a, b] \rightarrow \mathbb{R}$ is a bounded function whose set of discontinuities has measure zero, then f is Riemann-integrable.*

Proof. Define functions $F, H : [a, b] \rightarrow \mathbb{R}$ by the formulas

$$F(x) = \int_a^x f(t)dt, \quad H(x) = \int_x^a f(t)dt \quad (a \leq x \leq b).$$

(the *indefinite* upper and lower integrals of f). If f is continuous at a point $x \in (a, b)$, then F and H are differentiable at x , with $F'(x) = f(x) = H'(x)$,³ thus $(F - H)'(x) = 0$. In view of the hypothesis on f , $(F - H)' = 0$ a.e.; since F and H are absolutely continuous (even Lipschitz⁴), it follows that $F - H$ is constant (5.4.5). In particular,

$$F(b) - H(b) = F(a) - H(a) = 0 - 0;$$

³ First course, p. 148, 9.3.5.

⁴ First course, p. 146, 9.3.2.

the equality $F(b) = H(b)$ is precisely the conclusion of the lemma. {Note that the proof depends only on the concept of negligible set and requires none of the machinery of the Lebesgue integral.} \diamond

To prove the converse of the lemma, that the set of discontinuities of a Riemann-integrable function is negligible, we employ again the idea of semicontinuous approximation (cf. §5.9). Although the needed facts about semicontinuity will be applied here only for bounded functions on the metric space $[a, b]$, it is just as easy to develop them for extended-real-valued functions on a topological space (thereby making them available for application in other contexts).

5.13.2. *Definition.* Let X be a topological space, $f : X \rightarrow \overline{\mathbb{R}}$. With notations as in 5.5.15, we define functions $f_* : X \rightarrow \overline{\mathbb{R}}$ and $f^* : X \rightarrow \overline{\mathbb{R}}$ by the formulas

$$f_*(x) = \liminf_{t \rightarrow x} f(t) = \sup_V \left(\inf_{t \in V} f(t) \right),$$

$$f^*(x) = \limsup_{t \rightarrow x} f(t) = \inf_V \left(\sup_{t \in V} f(t) \right),$$

for every $x \in X$, where V runs over the set of all neighborhoods of x .

5.13.3. *Remarks.* (i) $f_* \leq f \leq f^*$ (5.5.16).

(ii) If $f \leq g$, then $f_* \leq g_*$ and $f^* \leq g^*$.

(iii) If f is bounded then so are f_* and f^* ; more precisely, if $m1 \leq f \leq M1$ ($m, M \in \mathbb{R}$) then also $m1 \leq f_* \leq f \leq f^* \leq M1$.

5.13.4. *Lemma.* With notations as in the above definition,

(1) f_* is lower semicontinuous;

(2) if $g : X \rightarrow \overline{\mathbb{R}}$ is lower semicontinuous and $g \leq f$, then $g \leq f_*$;

(3) f^* is upper semicontinuous;

(4) if $h : X \rightarrow \overline{\mathbb{R}}$ is upper semicontinuous and $f \leq h$, then $f^* \leq h$.

So to speak, f_* is the largest l.s.c. function $\leq f$, and f^* is the smallest u.s.c. function $\geq f$.

Proof. (1) Assuming $c \in X$ and $f_*(c) > r \in \mathbb{R}$, we have to show that $f_*(x) > r$ on some neighborhood W of c . By the definition of $f_*(c)$ as a least upper bound of infima, there exists a neighborhood W of c such that

$$\inf_{t \in W} f(t) > r;$$

such an infimum can only increase if the neighborhood is made smaller, so we can suppose that W is open. For every $x \in W$, W is a neighborhood of x , therefore

$$f_*(x) \geq \inf_{t \in W} f(t) > r.$$

(2) If $g \leq f$ and g is l.s.c., then $g = g_* \leq f_*$ (by 5.5.17 and the above remark (ii)).

The proofs of (3) and (4) are similar. \diamond

5.13.5. *Definition.* With notations as in 5.13.2, f_* is called the **lower semicontinuous regularization** of f , and f^* the **upper semicontinuous regularization** of f .⁵ (The terms ‘lower envelope’ and ‘upper envelope’ are also used.⁶)

5.13.6. *Lemma.* Every (upper or lower) semicontinuous function $f : [a, b] \rightarrow \mathbb{R}$ is a Borel function.

Proof. Suppose, for example, that f is lower semicontinuous. The claim is that for every Borel set B of \mathbb{R} , $f^{-1}(B)$ is a Borel set of \mathbb{R} contained in $[a, b]$; that is, in the notations of Example 4.7.6, $B \in \mathcal{B} \Rightarrow f^{-1}(B) \in \mathcal{B}_0$. It suffices to consider $B = (-\infty, r]$ (such intervals generate \mathcal{B}). Then $\mathbb{R} - B$ is open in \mathbb{R} , so by the lower semicontinuity of f we know that $f^{-1}(\mathbb{R} - B) = [a, b] - f^{-1}(B)$ is open in the metric space $[a, b]$ (5.5.12); thus $f^{-1}(B)$ is closed in $[a, b]$, hence in \mathbb{R} (because $[a, b]$ is closed in \mathbb{R}), therefore $f^{-1}(B) \in \mathcal{B}_0$. {The argument could be shortened a little, at the cost of discussing the ‘relative topology’ on a subset of \mathbb{R} (cf. §3.3, Exercise 7.)} \diamond

5.13.7. *Theorem. (Lebesgue’s criterion)* If $f : [a, b] \rightarrow \mathbb{R}$ is a bounded function, the following conditions are equivalent:

- (a) f is Riemann-integrable;
- (b) the set of points of discontinuity of f has Lebesgue measure zero.

Proof. (b) \Rightarrow (a): This is Lemma 5.13.1.

(a) \Rightarrow (b): Adding a constant to f , we can suppose that $f \geq 1$ on $[a, b]$. Let f_* and f^* be the lower and upper semicontinuous regularizations of f (5.13.5); in particular, $f_* \leq f \leq f^*$. Since f_* and f^* are also bounded (5.13.3, (iii)), it follows from the preceding lemma that they are Lebesgue-integrable.

Let $\sigma = \{a = a_0 < a_1 < \cdots < a_n = b\}$ be any subdivision of $[a, b]$ and adopt the notations preceding Lemma 5.13.1. Define “step” functions $g, h : [a, b] \rightarrow \mathbb{R}$ by the formulas

$$g = \sum_{i=1}^n m_i \varphi_{(a_{i-1}, a_i]}, \quad h = \sum_{i=1}^n M_i \varphi_{[a_{i-1}, a_i]},$$

where m_i, M_i are the bounds of f on $[a_{i-1}, a_i]$ and φ denotes characteristic function. Since $m_i > 0$ (recall that $f \geq 1$), each term of g

⁵ Cf. N. Bourbaki, *General topology. Vol. I* [Addison-Wesley, Reading, 1966], Ch. IV, §6, no. 2.

⁶ H. L. Royden, *Real analysis* [3rd. edn., Macmillan, 1988], p. 52.

is lower semicontinuous on $[a, b]$ (5.5.11), therefore so is g (5.5.21); similarly, h is upper semicontinuous on $[a, b]$. Moreover, it is clear from the positivity of f that $g \leq f \leq h$, therefore

$$g \leq f_* \leq f \leq f^* \leq h$$

by Lemma 5.13.4. It follows that

$$\int_a^b g \, d\lambda \leq \int_a^b f_* \, d\lambda \leq \int_a^b f^* \, d\lambda \leq \int_a^b h \, d\lambda,$$

that is,

$$s(\sigma) \leq \int_a^b f_* \, d\lambda \leq \int_a^b f^* \, d\lambda \leq S(\sigma),$$

where $s(\sigma)$ and $S(\sigma)$ are the lower and upper sums of f for σ ; since σ is arbitrary, we conclude that

$$(*) \quad \int_a^b f(t) \, dt \leq \int_a^b f_* \, d\lambda \leq \int_a^b f^* \, d\lambda \leq \int_a^b f(t) \, dt.$$

By hypothesis, f is Riemann-integrable, so the extremities of $(*)$ are equal, therefore

$$\int_a^b (f^* - f_*) \, d\lambda = 0,$$

and from $f^* - f_* \geq 0$ we infer further that $f^* - f_* = 0$ a.e. (4.4.21, (3)). Thus $f_* = f = f^*$ a.e.; if c is any point where $f_*(c) = f^*(c)$, then f is continuous at c (5.5.19). \diamond

A corollary of the proof of “(a) \Rightarrow (b)”:

5.13.8. Corollary. *If $f : [a, b] \rightarrow \mathbb{R}$ is Riemann-integrable, then f is also Lebesgue-integrable and*

$$\int_a^b f(t) \, dt = \int_a^b f \, d\lambda.$$

Proof. In the notations of $(*)$ of the preceding proof, $f = f^*$ a.e. and $f^* \in \mathcal{L}^1 = \mathcal{L}^1([a, b], \lambda)$, therefore $f \in \mathcal{L}^1$ and

$$\int_a^b f \, d\lambda = \int_a^b f^* \, d\lambda;$$

by equality in $(*)$,

$$\int_a^b f^* \, d\lambda = \int_a^b f(t) \, dt. \quad \diamond$$

A Riemann-integrable function is bounded and Lebesgue-integrable (5.13.8). In the Lebesgue theory, essentially bounded functions are on an

equal footing with bounded ones, and an essentially bounded measurable function on $[a, b]$ is Lebesgue-integrable; for such functions, there is a well-rounded “Fundamental theorem of calculus”:

5.13.9. Theorem. *The following conditions on a function $f : [a, b] \rightarrow \mathbb{R}$ are equivalent:*

(a) *f is essentially bounded and Lebesgue-measurable (hence Lebesgue-integrable on $[a, b]$);*

(b) *there exists a function $F : [a, b] \rightarrow \mathbb{R}$ satisfying a Lipschitz condition $|F(x) - F(y)| \leq K|x - y|$, such that $F' = f$ a.e.*

Any two such functions F differ by a constant, and

$$F(x) = F(a) + \int_a^x f \, d\lambda \quad \text{for all } x \in [a, b],$$

in particular

$$\int_a^b f \, d\lambda = F(b) - F(a).$$

Proof. (a) \Rightarrow (b): If $|f| \leq K$ a.e. and F is the indefinite integral of f , then $F' = f$ a.e. (5.9.3) and, if $a \leq x < y \leq b$, then

$$|F(y) - F(x)| = \left| \int_x^y f \, d\lambda \right| \leq \int_x^y |f| \, d\lambda \leq K|x - y|.$$

(b) \Rightarrow (a): Suppose K is a real number > 0 such that $|F(x) - F(y)| \leq K|x - y|$ for all x, y in $[a, b]$. It follows that F is absolutely continuous (5.1.10, (vi)), so there exists a Lebesgue-integrable function $g : [a, b] \rightarrow \mathbb{R}$ such that $F' = g$ a.e. (5.10.2); it is clear from the Lipschitz condition that $|F'| \leq K$, therefore $|g| \leq K$ a.e., thus g is essentially bounded. By assumption, $F' = f$ a.e., therefore $f = g$ a.e.; thus f is also Lebesgue-measurable and essentially bounded.

The last assertions of the theorem are immediate from 5.10.1. \diamond

Inspection of the above proof yields the following characterization of the ‘primitives’ that occur in the theorem:

5.13.10. Corollary. *For a function $F : [a, b] \rightarrow \mathbb{R}$, the following conditions are equivalent:*

(a) *F is the indefinite integral of an essentially bounded, Lebesgue-measurable function $f : [a, b] \rightarrow \mathbb{R}$;*

(b) *$F(a) = 0$ and F satisfies a Lipschitz condition $|F(x) - F(y)| \leq K|x - y|$.*

Is there a “Fundamental theorem of calculus” for Riemann-integrable functions? The best we can squeeze out of the theorems of this section is the following:

Let $f : [a, b] \rightarrow \mathbb{R}$ be Riemann-integrable. Then

(1) *the indefinite integral*

$$F(x) = \int_a^x f(t)dt \quad (a \leq x \leq b)$$

of f is absolutely continuous (even Lipschitz) and satisfies $F' = f$ a.e.;

(2) if also $G : [a, b] \rightarrow \mathbb{R}$ is absolutely continuous and $G' = f$ a.e., then G differs from F by a constant, therefore

$$G(x) = G(a) + \int_a^x f(t)dt \quad \text{for all } x \in [a, b].$$

The only thing in view resembling an ‘integral-free’ characterization of ‘primitives’ for the Riemann theory is condition (b) of the following near-tautology:

For a function $F : [a, b] \rightarrow \mathbb{R}$, the following conditions are equivalent:

- (a) F is the indefinite integral of a Riemann-integrable function on $[a, b]$;
- (b) $F(a) = 0$, F is absolutely continuous, and there exists a bounded function $f : [a, b] \rightarrow \mathbb{R}$, with negligible set of discontinuities, such that $F' = f$ a.e.

With notations as in (b), f is Riemann-integrable and F is its indefinite integral.

The “Fundamental theorem of calculus” that the Riemann integral would like to enjoy seems to be preempted by the essentially bounded measurable functions (5.13.9 and 5.13.10); what is missing in the Riemann case is a condition on F , stronger than Lipschitz, that does not give the show away like the above condition (b).

Exercises

1. (i) The function $F : [0, 1] \rightarrow \mathbb{R}$ defined by $F(x) = (1 - x^2)^{1/2}$ is continuous on $[0, 1]$ and differentiable on $(0, 1)$, but it is not the primitive of a continuous function $f : [0, 1] \rightarrow \mathbb{R}$. {Hint: F' is unbounded.}

(ii) Let $g : [0, 1] \rightarrow \mathbb{R}$ be the (Riemann-integrable) function defined by

$$g(x) = \begin{cases} \sin(1/x) & \text{for } x \in (0, 1] \\ 0 & \text{for } x = 0 \end{cases}$$

and let $F : [0, 1] \rightarrow \mathbb{R}$ be the indefinite integral of g ,

$$F(x) = \int_0^x g(t)dt \quad (0 \leq x \leq 1).$$

Then F is continuous on $[0, 1]$, differentiable on $(0, 1)$, and F' is bounded, but F is not a primitive of a continuous function $f : [0, 1] \rightarrow \mathbb{R}$. {Hint: $g(0+)$ does not exist.}

2. Let $f : [a, b] \rightarrow \mathbb{R}$ be Riemann-integrable, F its indefinite integral, $c \in [a, b)$.

(i) If f has a right limit L at c , then F is right-differentiable at c and $F'_r(c) = L$.

{Hint: Redefining f at c , one can suppose that f is right-continuous at c .}

(ii) The converse of (i) is false. For example, if $f : [0, 2] \rightarrow \mathbb{R}$ is the characteristic function of the set

$$A = \{1 - 1/n : n \in \mathbb{P}\} \cup \{1 + 1/n : n \in \mathbb{P}\},$$

then f is Riemann-integrable, its indefinite integral F is identically zero, $F'(1) = f(1)$ but neither $f(1+)$ nor $f(1-)$ exist.

3. If $f : [a, b] \rightarrow \mathbb{R}$ is any bounded function, then its set of discontinuities is the union of a sequence of closed sets (i.e., is an F_σ set in \mathbb{R}).

{Hint: With notations as in Definition 5.13.5, argue that the set $\{x \in [a, b] : f^*(x) - f_*(x) \geq 1/n\}$ is a closed set (cf. 5.5.13).}

CHAPTER 6

Function Spaces

- §6.1. Compact metric spaces
- §6.2. Uniform convergence, iterated limits theorem
- §6.3. Complete metric spaces
- §6.4. L^1
- §6.5. Real and complex measures
- §6.6. L^∞
- §6.7. L^p ($1 < p < +\infty$)
- §6.8. $\mathcal{C}(X)$
- §6.9. Stone-Weierstrass approximation theorem

The emphasis of the present chapter is on metric spaces whose elements are functions (or equivalence classes of functions). The main examples arise in topological or measure-theoretic contexts; the first three sections prepare the way with the necessary topics in topology and metric spaces.

6.1. Compact Metric Spaces

The concept of compactness to be defined shortly is motivated by the following property of a closed interval:

6.1.1. Theorem. (Heine–Borel)¹ *Let $[a, b]$ be a closed interval in \mathbb{R} and suppose \mathcal{C} is a class of open sets in \mathbb{R} such that each point of $[a, b]$ belongs to at least one of the sets of \mathcal{C} , briefly $[a, b] \subset \bigcup \mathcal{C}$. Then*

$$[a, b] \subset U_1 \cup \dots \cup U_n$$

for a suitable finite list U_1, \dots, U_n of sets in \mathcal{C} .

Proof. Let S be the set of all points x of $[a, b]$ such that the closed interval $[a, x]$ is contained in the union of a finite number of sets in \mathcal{C} . Since $[a, a] = \{a\} \subset U$ for some $U \in \mathcal{C}$, we have $a \in S$; our problem is to show that $b \in S$.

¹ Heinrich Eduard Heine (1821–1881), Émile Borel (1871–1956).

The set S is nonempty and bounded; let $M = \sup S$. Obviously $a \leq M \leq b$. We will show that $b \in S$ by proving that (1) $M \in S$, and (2) $M = b$.

(1) Since $M \in [a, b]$, by assumption there exists a set $V \in \mathcal{C}$ such that $M \in V$; since V is open, we have $[M - \epsilon, M + \epsilon] \subset V$ for a suitable $\epsilon > 0$. We note for use in the proof of (2) that ϵ can be taken to be as small as we like.

Choose $x \in S$ so that $M - \epsilon < x \leq M$ (possible because M is the least upper bound of S). From $x \in S$ we know that the interval $[a, x]$ is contained in the union of finitely many sets in \mathcal{C} , say

$$[a, x] \subset U_1 \cup \dots \cup U_r;$$

since $[x, M] \subset [M - \epsilon, M + \epsilon] \subset V$, it follows that

$$[a, M] = [a, x] \cup [x, M] \subset U_1 \cup \dots \cup U_r \cup V,$$

therefore $M \in S$. Moreover, since $[M, M + \epsilon] \subset V$ we have in fact

$$[a, M + \epsilon] \subset U_1 \cup \dots \cup U_r \cup V.$$

It follows that $M + \epsilon > b$; for, the alternative $M + \epsilon \leq b$ would imply (by the preceding inclusion) that $M + \epsilon \in S$, contrary to the fact that every element of S is $\leq M$.

(2) The foregoing argument shows that $b < M + \epsilon$ for arbitrarily small ϵ , therefore $b \leq M$; already $M \leq b$, so $b = M \in S$. \diamond

For use in more general situations later on, we separate out several concepts involved in the Heine–Borel theorem in a form applicable in general topological spaces (not necessarily derived from a metric):

6.1.2. *Definition.* Let X be a topological space and let A be a subset of X . A set \mathcal{C} of subsets of X is said to be a **covering** of A if $A \subset \bigcup \mathcal{C}$, that is, if each point of A belongs to at least one of the sets in \mathcal{C} ; if $\mathcal{D} \subset \mathcal{C}$ and \mathcal{D} is also a covering of A , then \mathcal{D} is said to be a **subcovering** of \mathcal{C} . (The language is regrettably awkward; \mathcal{D} is a subset of \mathcal{C} but it is A that gets covered.) A covering \mathcal{C} of A consisting of a finite number of sets is called a **finite covering** of A ; a covering \mathcal{C} of A whose elements are open sets in X is called an **open covering** of A .

Expressed in the foregoing language, the Heine–Borel theorem asserts that (in the topological space \mathbb{R}) every open covering of a closed interval $[a, b]$ has a finite subcovering. This prompts the next definition:

6.1.3. *Definition.* A subset A of a topological space X is said to be **quasicompact** if every open covering of A has a finite subcovering; expressed in the notation of indexed families, this means that if $(U_i)_{i \in I}$ is a family of open sets in X such that $A \subset \bigcup_{i \in I} U_i$, then there exists a finite subset J of I such that $A \subset \bigcup_{i \in J} U_i$. If X is a quasicompact subset of itself then it is called a **quasicompact space**.

Reformulated in terms of closed sets, quasicompactness has the flavor of an ‘induction principle’ (from finite to infinite):

6.1.4. Theorem. *The following conditions on a topological space X are equivalent:*

- (a) X is quasicompact;
- (b) if $(F_i)_{i \in I}$ is a family of closed sets in X such that $\bigcap_{i \in J} F_i \neq \emptyset$ for every finite subset J of I , then $\bigcap_{i \in I} F_i \neq \emptyset$.

Proof. Stated contrapositively, condition (b) says that, for a family $(F_i)_{i \in I}$ of closed sets,

$$\bigcap_{i \in I} F_i = \emptyset \quad \Rightarrow \quad \bigcap_{i \in J} F_i = \emptyset \quad \text{for some finite } J \subset I,$$

in other words,

$$\bigcup_{i \in I} \mathcal{C}F_i = X \quad \Rightarrow \quad \bigcup_{i \in J} \mathcal{C}F_i = X \quad \text{for some finite } J \subset I.$$

In view of the duality between closed sets and open sets (3.3.1), it is clear that (b) is equivalent to the assertion that every open covering of X has a finite subcovering, which is also the meaning of (a). \diamond

The hypothesis in condition (b) can be expressed by saying that the family $(F_i)_{i \in I}$ has the *finite intersection property* (every finite subfamily has nonempty intersection); condition (b) then says that every family of closed sets with the finite intersection property has nonempty intersection.

6.1.5. Corollary. *If X is a quasicompact topological space and if (F_n) is a sequence of nonempty closed sets in X such that $F_1 \supset F_2 \supset F_3 \supset \dots$, then $\bigcap_{n=1}^{\infty} F_n \neq \emptyset$.*

Proof. It is obvious that the family $(F_n)_{n \in \mathbb{P}}$ has the finite intersection property. \diamond

The definition of compactness requires quasicompactness and one extra condition:

6.1.6. Definition. A topological space is said to be **separated** (or to be a Hausdorff space²) if, for every pair of distinct points x and y of the space, there exist open sets U and V such that $x \in U$, $y \in V$ and $U \cap V = \emptyset$ (so to speak, distinct points can be separated by means of disjoint open sets—or, equivalently, by means of disjoint neighborhoods of the points). A topological space is said to be **compact** if it is both quasicompact and separated.

6.1.7. Remarks. (i) Every metric space (X, d) is separated for the topology \mathcal{O}_d derived from its metric. {Proof: If $x \neq y$ and if $r = \frac{1}{2}d(x, y)$,

² After Felix Hausdorff (1868–1942).

then the open balls $U = U_r(x)$ and $V = U_r(y)$ are disjoint neighborhoods of x and y respectively; for, the existence of a point $z \in U \cap V$ would imply that $d(x, y) \leq d(x, z) + d(z, y) < r + r = d(x, y)$.} Thus, for a metric space, the concepts of compactness and quasicompactness coincide.

(ii) A quasicompact space need not be compact (consider a two-point set equipped with the trivial topology (3.3.2)).

(iii) Let (X, \mathcal{O}) be a topological space and let A be a subset of X . The class

$$\mathcal{O} \cap A = \{U \cap A : U \in \mathcal{O}\}$$

of subsets of A is easily seen to be a topology on A ; it is called the *relative topology* on A induced by \mathcal{O} (cf. §3.3, Exercise 7). One also writes $\mathcal{O}_A = \mathcal{O} \cap A$, and (A, \mathcal{O}_A) is called a (topological) *subspace* of (X, \mathcal{O}) . If \mathcal{C} is an open covering of A in the sense of 6.1.2, then $\mathcal{C} \cap A = \{U \cap A : U \in \mathcal{C}\}$ is a class of open subsets of A (for the relative topology) whose union is A ; it follows easily that A is a quasicompact subset of X (in the sense of 6.1.3) if and only if A is a quasicompact space for the relative topology.

(iv) With notations as in (iii), if X is separated then A is separated for the relative topology (if U and V are disjoint, then so are $U \cap A$ and $V \cap A$).

6.1.8. *Definition.* A subset A of a topological space (X, \mathcal{O}) is said to be **compact** if, for the relative topology induced by \mathcal{O} , A is a compact topological space, that is, if (A, \mathcal{O}_A) is a compact space in the sense of Definition 6.1.6.

6.1.9. *Examples.* (1) Let X be a separated topological space (for example, a metric space) and let A be a subset of X . In view of (iii) and (iv) of 6.1.7, A is a compact subset of X if and only if it is a quasicompact subset of X . In particular, the Heine–Borel theorem asserts that every closed interval $[a, b]$ is a compact subset of \mathbb{R} ; in other words (cf. Exercise 2), $[a, b]$ is a compact metric space for the usual metric $(x, y) \mapsto |x - y|$.

(2) If (x_n) is a convergent sequence in a metric space X , say $x_n \rightarrow x$, then the set

$$A = \{x\} \cup \{x_n : n = 1, 2, 3, \dots\}$$

is a compact subset of X .

{Hint: An open set containing x contains all but finitely many of the x_n (cf. 3.2.19).}

The Weierstrass–Bolzano theorem (cf. 1.16.11) states that every bounded sequence in \mathbb{R} has a convergent subsequence. In particular, every sequence in a closed interval $[a, b]$ of \mathbb{R} has a convergent subsequence, whose limit is in $[a, b]$ because $[a, b]$ is a closed subset of \mathbb{R} .

6.1.10. *Definition.* A metric space is said to have the *Weierstrass-Bolzano property* if every sequence in the space has a convergent subsequence.

The main goal of this section is to prove that *a metric space is compact (for the topology derived from its metric) if and only if it has the Weierstrass-Bolzano property.* Some of the most important metric space concepts figure in the proof (total boundedness, separability, completeness); the proof is organized in a series of lemmas, interspersed with the definitions of these concepts and some examples. Half of the equivalence is disposed of by the first lemma:

6.1.11. *Lemma.* If (X, d) is a compact metric space, then every sequence in X has a convergent subsequence.

Proof. Let (x_n) be a sequence in X . For each index n , let

$$A_n = \{x_k : k > n\}.$$

The sets A_n are nonempty and $A_1 \supset A_2 \supset A_3 \supset \dots$. Since the closure operation preserves inclusion (3.3.16), we have

$$\overline{A_1} \supset \overline{A_2} \supset \overline{A_3} \supset \dots;$$

by compactness, the intersection of the sets $\overline{A_n}$ is nonempty, say $x \in \bigcap_{n=1}^{\infty} \overline{A_n}$. We will show that x is the limit of a suitable subsequence of (x_n) .

Since x is adherent to A_1 , there exists an index $n_1 > 1$ such that $d(x_{n_1}, x) < 1$; then, since x is adherent to A_{n_1} , there exists an index $n_2 > n_1$ such that $d(x_{n_2}, x) < 1/2$. Continuing recursively, we obtain a sequence of indices $n_1 < n_2 < n_3 < \dots$ such that $d(x_{n_k}, x) < 1/k$, thus (x_{n_k}) is a subsequence of (x_n) with $d(x_{n_k}, x) \rightarrow 0$ as $k \rightarrow \infty$. \diamond

This proves that every compact metric space has the Weierstrass-Bolzano property. Before proving the reverse implication, let us note a property of compactness that motivates the next definition:

6.1.12. *Remark.* If (X, d) is a compact metric space then, for every $\epsilon > 0$, there exists a finite list of points y_1, \dots, y_r in X such that each point of X is within ϵ of at least one of the y_i , that is,

$$X = \bigcup_{i=1}^r U_{\epsilon}(y_i)$$

(of course r , and the points y_1, \dots, y_r , will in general depend on ϵ). {Proof: The open balls $U_{\epsilon}(y)$, $y \in X$, constitute an open covering of X ; pass to a finite subcovering.}

6.1.13. *Definition.* Let (X, d) be a metric space and let $\epsilon > 0$. An ϵ -net in X is a finite subset F of X such that

$$X = \bigcup_{y \in F} U_\epsilon(y).$$

Thus, if $F = \{y_1, \dots, y_r\}$, then every point of X is within ϵ of a least one of the points y_i . The metric space (X, d) is said to be **totally bounded** if it has an ϵ -net for every $\epsilon > 0$ (it clearly suffices that there exist a $\frac{1}{n}$ -net for every positive integer n).

For example, every compact metric space is totally bounded (6.1.12), but the converse is false (cf. Exercise 5).

If $x, y \in U_r(a)$ then $d(x, y) \leq 2r$ by the triangle inequality. This prompts the next definition:

6.1.14. *Definition.* Let (X, d) be a metric space, A a nonempty subset of X . We say that A has **finite diameter** if there exists a real number $K \geq 0$ such that

$$d(x, y) \leq K \text{ for all } x, y \in A;$$

more precisely, the **diameter** of such a set, denoted $\text{diam } A$, is defined to be the infimum of all such K ,

$$\text{diam } A = \inf\{K : d(x, y) \leq K \text{ for all } x, y \in A\},$$

and it is clear from the definition of suprema that $\text{diam } A = \sup\{d(x, y) : x, y \in A\}$.

6.1.15. *Examples.* (i) In a metric space, every ball (open or closed) of radius r has diameter $\leq 2r$, and every subset of finite diameter is contained in some ball.

(ii) In a discrete metric space (3.1.7) every open ball of radius $r \leq 1$ has diameter 0.

(iii) A metric space (X, d) is totally bounded if and only if, for every $\epsilon > 0$, X is the union of finitely many sets of diameter $\leq \epsilon$.

The next definition is a generalization to metric spaces of a concept familiar from elementary analysis:

6.1.16. *Definition.* A sequence (x_n) in a metric space (X, d) is said to be a **Cauchy sequence** if $d(x_m, x_n) \rightarrow 0$ as $m, n \rightarrow \infty$, in the following sense: for every $\epsilon > 0$ there exists an index N such that $d(x_m, x_n) < \epsilon$ for all $m, n \geq N$.

Every convergent sequence is Cauchy; for, if $d(x_n, x) \rightarrow 0$ then $d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) < \epsilon$ provided that $d(x_m, x) < \epsilon/2$ and $d(x_n, x) < \epsilon/2$. The converse is false; for example, in the open interval $X = (0, +\infty)$ equipped with the usual metric $d(x, y) = |x - y|$, the

sequence $x_n = 1/n$ is Cauchy (because it is convergent in \mathbb{R}) but has no limit in X .

6.1.17. Lemma. *If (X, d) is a metric space in which every sequence has a Cauchy subsequence, then the space is totally bounded.*

Proof. (The converse is also true—see 6.1.24 below.) We argue contrapositively: assuming that X is *not* totally bounded, let us construct a sequence (x_n) in X that has *no* Cauchy subsequence. By assumption, there exists an $\epsilon > 0$ for which no ϵ -net exists; that is, every finite subset of X fails to be an ϵ -net. Thus, for every finite subset F of X , there exists a point $x \in X$ such that $d(x, y) \geq \epsilon$ for all $y \in F$. The construction of (x_n) proceeds as follows.

Choose any point x_1 in X . Since $\{x_1\}$ is not an ϵ -net, there exists a point x_2 such that $d(x_2, x_1) \geq \epsilon$. Since $\{x_1, x_2\}$ is not an ϵ -net, there exists a point x_3 such that $d(x_3, x_1) \geq \epsilon$ and $d(x_3, x_2) \geq \epsilon$. Continuing in the obvious recursive way, we obtain a sequence (x_n) such that $d(x_m, x_n) \geq \epsilon$ whenever $m \neq n$, a sequence that can have no Cauchy subsequence. \diamond

6.1.18. Definition. A metric space is said to be **separable** if it has a countable dense subset, that is, a countable subset A such that $\overline{A} = X$.

For example, the real number field \mathbb{R} equipped with the usual metric is separable, with the rational field \mathbb{Q} as a countable (1.10.10) dense subset (1.8.25). An uncountable discrete metric space is not separable.

6.1.19. Lemma. *Every totally bounded metric space (X, d) is separable.*

Proof. For each positive integer n , let F_n be a $\frac{1}{n}$ -net in X . The set $A = \bigcup_{n=1}^{\infty} F_n$ is countable; we will show that it is dense in X . It suffices to show that every open ball $U_r(x)$ has nonempty intersection with A . Choose n so that $\frac{1}{n} < r$. Since F_n is a $\frac{1}{n}$ -net, there exists a point $y \in F_n$ such that $d(x, y) < 1/n$; then $y \in U_r(x)$ (because $\frac{1}{n} < r$) and $y \in A$ (because $F_n \subset A$), thus $U_r(x) \cap A \neq \emptyset$. \diamond

6.1.20. Definition. Let X be a topological space and let \mathcal{B} be a set of open sets in X ; \mathcal{B} is said to be a **base** for the topology of X (or for the open sets of X) if every open set is a union of sets in \mathcal{B} ; equivalently,

$$U \text{ open, } x \in U \Rightarrow \exists V \in \mathcal{B} \ni x \in V \subset U.$$

So to speak, the sets of \mathcal{B} ‘pry into every neighborhood’: between any point x and any of its neighborhoods, one can interpolate one of the sets of the base \mathcal{B} .

6.1.21. Lemma. *Every separable metric space has a countable base for the topology derived from the metric.*

Proof. Let $A = \{a_k : k = 1, 2, 3, \dots\}$ be a countable dense subset of the space (6.1.18) and let

$$\mathcal{B} = \{U_{1/n}(a_k) : n, k \in \mathbb{P}\}$$

be the set of all open balls, centered at the a_k , with radii $1/n$ ($n \in \mathbb{P}$). Clearly \mathcal{B} is a countable set (cf. 1.10.8) of open sets; we will show that it is a base for the topology.

Let U be an open set and let $x \in U$; we are to interpolate a set of \mathcal{B} between x and U . Choose $r > 0$ so that $U_r(x) \subset U$, let n be a positive integer such that $\frac{1}{n} < \frac{r}{2}$, and let k be an index such that $d(a_k, x) < 1/n$ (possible because A is dense). Then

$$x \in U_{1/n}(a_k) \subset U;$$

for, if $y \in U_{1/n}(a_k)$ then

$$d(y, x) \leq d(y, a_k) + d(a_k, x) < \frac{1}{n} + \frac{1}{n} = \frac{2}{n} < r,$$

thus $y \in U_r(x) \subset U$. \diamond

The property of having a countable base in fact characterizes the separable metric spaces (Exercise 7).

Proving compactness entails finding finite subcoverings; finding a countable subcovering, which can be a valuable intermediate step, is available in every space with a countable base for open sets:

6.1.22. Lemma. (Lindelöf's theorem) *In a topological space with a countable base for the open sets, every open covering of the space has a countable subcovering.*

Proof. Let \mathcal{B} be a countable base for the open sets of the topological space X and let \mathcal{U} be any open covering of X ; we seek a countable subcovering \mathcal{U}_0 of \mathcal{U} . Let

$$\mathcal{B}_0 = \{V \in \mathcal{B} : V \subset U \text{ for some } U \in \mathcal{U}\};$$

since $\mathcal{B}_0 \subset \mathcal{B}$, \mathcal{B}_0 is countable (1.10.2), say $\mathcal{B}_0 = \{V_n : n \in \mathbb{P}\}$.

For each positive integer n , choose a set $U_n \in \mathcal{U}$ with $V_n \subset U_n$ (possible by the definition of \mathcal{B}_0) and let $\mathcal{U}_0 = \{U_n : n \in \mathbb{P}\}$; \mathcal{U}_0 is a countable subset of \mathcal{U} , and we need only show that it is a covering of X .

Let $x \in X$; we seek an index n such that $x \in U_n$. Choose $U \in \mathcal{U}$ with $x \in U$ (\mathcal{U} is a covering of X) and let $V \in \mathcal{B}$ with $x \in V \subset U$ (\mathcal{B} is a base for the topology); then $V \in \mathcal{B}_0$ by the definition of \mathcal{B}_0 , thus $V = V_n$ for some n , and finally $x \in V_n \subset U_n$. \diamond

We can now characterize compact metric spaces as the metric spaces having the Weierstrass-Bolzano property:

6.1.23. Theorem. *The following conditions on a metric space (X, d) are equivalent:*

- (a) X is compact (for the topology \mathcal{O}_d derived from the metric d);
 (b) every sequence in X has a convergent subsequence.

Proof. (a) \Rightarrow (b): This is Lemma 6.1.11.

(b) \Rightarrow (a): Since convergent sequences are Cauchy, we know that every sequence in X has a Cauchy subsequence, therefore X is totally bounded (6.1.17), hence separable (6.1.19), hence there is a countable base for the open sets (6.1.21).

Given any open covering \mathcal{U} of X , we seek a finite subcovering. By Lindelöf's theorem (6.1.22) we can suppose that \mathcal{U} is countable, say $\mathcal{U} = \{U_n : n \in \mathbb{P}\}$. For every positive integer n , let

$$V_n = U_1 \cup \dots \cup U_n;$$

we know that $V_n \uparrow X$ and it will suffice to show that $V_n = X$ for some n . Assume to the contrary that no such n exists, that is, $X - V_n \neq \emptyset$ for all n . For each n select a point $x_n \in X - V_n$. By hypothesis, the sequence (x_n) has a convergent subsequence, say $x_{n_k} \rightarrow x$. By monotonicity, $V_{n_k} \uparrow X$, so $x \in V_{n_j}$ for some j ; since V_{n_j} is open, $x_{n_k} \in V_{n_j}$ ultimately. Choose any k such that $k > j$ and $x_{n_k} \in V_{n_j}$. Then

$$x_{n_k} \in V_{n_j} \subset V_{n_k},$$

contrary to $x_{n_k} \in X - V_{n_k}$. \diamond

The totally bounded spaces are characterized by a 'Cauchy' variant of the Weierstrass-Bolzano property:

6.1.24. Theorem. *The following conditions on a metric space (X, d) are equivalent:*

- (a) X is totally bounded;
 (b) every sequence in X has a Cauchy subsequence.

Proof. (b) \Rightarrow (a): This is Lemma 6.1.17.

(a) \Rightarrow (b): Let (x_n) be a sequence in X ; assuming X is totally bounded, we seek a Cauchy subsequence (x_{n_k}) .

Given any $\epsilon > 0$, X is expressible as a finite union of open balls of radius $\epsilon/2$, hence of diameter $\leq \epsilon$. It follows that every subset A of X is the union of finitely many sets of diameter $\leq \epsilon$; if, moreover, $x_n \in A$ for infinitely many n , then one of the terms of such a union must contain x_n for infinitely many n . Summarizing, if A is a subset of X containing x_n for infinitely many n , then, given any $\epsilon > 0$, A has a subset B of diameter $\leq \epsilon$ that contains x_n for infinitely many n .

We now construct a sequence (A_k) of subsets of X such that the k 'th term x_{n_k} of the desired Cauchy subsequence will be drawn from A_k . By the preceding paragraph (with $A = X$ and $\epsilon = 1$) there exists a subset A_1 of X such that $\text{diam } A_1 \leq 1$ and $x_n \in A_1$ frequently. Similarly (with $A = A_1$ and $\epsilon = 1/2$) there exists a subset A_2 of A_1 such that $\text{diam } A_2 \leq 1/2$ and $x_n \in A_2$ frequently. One continues recursively

in the obvious way, obtaining a sequence $A_1 \supset A_2 \supset A_3 \supset \dots$ such that $\text{diam } A_k \leq 1/k$ and such that, for each k , $x_n \in A_k$ for infinitely many n .

The desired subsequence (x_{n_k}) of (x_n) is now constructed as follows. Choose any index n_1 such that $x_{n_1} \in A_1$. Then choose any index $n_2 > n_1$ such that $x_{n_2} \in A_2$. Recursively, choose $n_k > n_{k-1}$ such that $x_{n_k} \in A_k$. The resulting subsequence (x_{n_k}) is Cauchy. For, given any $\epsilon > 0$, there is an index k such that $\frac{1}{n_k} < \epsilon$; for every pair of indices $i, j \geq k$, we have

$$x_{n_i} \in A_{n_i} \subset A_{n_k} \quad \text{and} \quad x_{n_j} \in A_{n_j} \subset A_{n_k},$$

therefore $d(x_{n_i}, x_{n_j}) \leq \text{diam } A_{n_k} \leq 1/n_k < \epsilon$. \diamond

As remarked following 6.1.13, 'compact \Rightarrow totally bounded'. What can be added to total boundedness to convert the implication \Rightarrow into an equivalence \Leftrightarrow ? 'Completeness' does the job:

6.1.25. *Definition.* A metric space is said to be **complete** if every Cauchy sequence in the space is convergent to a point in the space.

The classical example of a complete metric space: the real number field \mathbb{R} , equipped with the usual metric $(x, y) \mapsto |x - y|$ (cf. 1.8.26). Other examples are given in the exercises, and complete metric spaces are studied in greater depth in Section 3 of this chapter.

6.1.26. *Theorem.* The following conditions on a metric space (X, d) are equivalent:

- (a) X is compact (for the topology \mathcal{O}_d derived from the metric d);
- (b) (X, d) is complete and totally bounded.

Proof. (a) \Rightarrow (b): By the remark following 6.1.13, X is totally bounded; we are to show that every Cauchy sequence (x_n) is convergent. By 6.1.11, (x_n) has a convergent subsequence, say $x_{n_k} \rightarrow x$, and it will suffice to show that $x_n \rightarrow x$. Given any $\epsilon > 0$, choose an index N such that $d(x_n, x_m) < \epsilon/2$ for all $n, m \geq N$, then choose an index k such that both $d(x_{n_k}, x) < \epsilon/2$ and $n_k \geq N$; then

$$d(x_n, x) \leq d(x_n, x_{n_k}) + d(x_{n_k}, x) < \frac{\epsilon}{2} + \frac{\epsilon}{2}$$

for all $n \geq N$.

(b) \Rightarrow (a): By Theorem 6.1.23, we need only show that every sequence (x_n) has a convergent subsequence. By total boundedness, (x_n) has a Cauchy subsequence (6.1.24) which, by completeness, is convergent. \diamond

Compactness and completeness figure prominently in the rest of the book; we record here a theorem concerning each of these concepts, both for application in the next section. The first is a mapping property of quasi-compactness:

6.1.27. Theorem. If $f: X \rightarrow Y$ is a continuous mapping between topological spaces X and Y and if A is a quasicompact subset of X , then its image $f(A)$ is a quasicompact subset of Y .

Proof. Assuming $(V_i)_{i \in I}$ is a family of open sets in Y with $f(A) \subset \bigcup_{i \in I} V_i$, we seek a finite subset J of I such that $f(A) \subset \bigcup_{j \in J} V_j$. We have

$$A \subset f^{-1}\left(\bigcup_{i \in I} V_i\right) = \bigcup_{i \in I} f^{-1}(V_i);$$

since A is quasicompact and the $f^{-1}(V_i)$ are open sets in X (3.4.5), there exists a finite subset J of I such that $A \subset \bigcup_{j \in J} f^{-1}(V_j)$, in other words $f(A) \subset \bigcup_{j \in J} V_j$. \diamond

6.1.28. Remark. Every nonempty subset A of a metric space (X, d) can itself be regarded as a metric space simply by restricting the given distance function d to pairs of points of A (cf. Exercise 2). Equipped with the restricted metric $d_A = d|_{A \times A}$, A is called a *metric subspace* of X .

6.1.29. Theorem. Let (X, d) be a metric space and let A be a nonempty subset of X , regarded as a metric subspace of X (6.1.28).

(1) If A is a complete metric subspace of X , then A is a closed set in X .

(2) If X is a complete metric space and A is a closed subset of X , then A is a complete metric subspace of X .

Proof. (1) If $a_n \in A$ and $d(a_n, x) \rightarrow 0$, then (a_n) is Cauchy in X , hence Cauchy in A , hence convergent to some point a of A , $d(a_n, a) \rightarrow 0$; then $x = a \in A$ by the uniqueness of limits (3.2.1), thus A is a closed set in X (3.2.5).

(2) Assuming A is a closed set in X , suppose that (a_n) is a Cauchy sequence in A , that is, $d(a_m, a_n) \rightarrow 0$ as $m, n \rightarrow \infty$; then (a_n) is also Cauchy in X , so by hypothesis there exists a point $x \in X$ with $d(a_n, x) \rightarrow 0$, and, since A is closed, necessarily $x \in A$. Thus, every Cauchy sequence in A is convergent in A . \diamond

Exercises

1. Let X be an infinite set and declare a subset U of X to be *open* if either $U = \emptyset$ or $X - U$ is finite. This defines a topology on X for which X is (1) quasicompact, but (2) not separated.

{Hint: (1) If $(U_i)_{i \in I}$ is an open covering of X and if j is an index such that $U_j \neq \emptyset$, then $X - U_j$ is covered by finitely many of the U_i . (2) $\mathcal{C}(U \cap V) = \mathcal{C}U \cup \mathcal{C}V$; infer that if U and V are nonempty open sets, then $U \cap V$ is infinite.}

2. If (X, d) is a metric space, A is a nonempty subset of X , and d_A is the restriction of d to $A \times A$, then d_A is a metric on A and the topology \mathcal{O}_{d_A} on A derived from d_A coincides with the relative topology $\mathcal{O}_d \cap A$ induced by \mathcal{O}_d (cf. §3.3, Exercise 7).

3. A subset of \mathbb{R} is compact if and only if it is bounded and closed.

4. The set $\overline{\mathbb{R}}$ of extended real numbers is a compact space for the topology derived from the metric defined in 3.3.17.

{Hint: If U and V are neighborhoods of $-\infty$ and $+\infty$ respectively, then $\mathbb{R} - (U \cup V)$ is contained in a closed interval $[a, b]$.}

5. The open interval $X = (0, 1)$, equipped with the usual metric $d(x, y) = |x - y|$ is totally bounded (for each integer $m > 1$, the points $\frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}$ form a $\frac{1}{m}$ -net) but not compact (the sequence $x_n = \frac{1}{n}$ has no convergent subsequence).

6. If X is an infinite set and d is the discrete metric on X (3.1.7) then (X, d) is complete but not totally bounded.

7. If a topological space (X, \mathcal{O}) has a countable base \mathcal{B} for the open sets, then it has a countable dense subset A (form A by choosing a point from each set in \mathcal{B}). Thus a metric space is separable if and only if it has a countable base for the open sets.

8. The real number field \mathbb{R} is complete for the usual metric (1.8.26). There exists a metric d on \mathbb{R} , equivalent (cf. 3.3.3) to the usual metric, such that the metric space (\mathbb{R}, d) is not complete.

{Hint: Let $f: \mathbb{R} \rightarrow (-1, 1)$ be a bijection such that both f and f^{-1} are continuous for the usual topologies, that is, for $x_n, x \in \mathbb{R}$,

$$|x_n - x| \rightarrow 0 \quad \Leftrightarrow \quad |f(x_n) - f(x)| \rightarrow 0.$$

(For example, the function

$$f(x) = \begin{cases} \frac{x}{1+x} & \text{for } x \geq 0 \\ \frac{x}{1-x} & \text{for } x < 0 \end{cases}$$

meets the requirements.) Define $d(x, y) = |f(x) - f(y)|$ for $x, y \in \mathbb{R}$. Then d is a metric on \mathbb{R} , equivalent to the usual metric, such that (\mathbb{R}, d) is not complete (the sequence $x_n = n$ is Cauchy for d , but not convergent).}

9. Let (X, d) be a metric space and let $D = d/(1 + d)$ (cf. Example 3.1.8). Recall that a sequence (x_n) in X is convergent for d if and only if it is convergent for D (3.2.2, (iii)). Prove:

- (i) (x_n) is Cauchy for $d \Leftrightarrow (x_n)$ is Cauchy for D .
- (ii) (X, d) is totally bounded $\Leftrightarrow (X, D)$ is totally bounded.
- (iii) (X, d) is complete $\Leftrightarrow (X, D)$ is complete.

(iv) (X, \mathcal{O}_d) is compact $\Leftrightarrow (X, \mathcal{O}_D)$ is compact.

{Hint: (i) $d = D/(1 - D)$. (ii) Cf. Theorem 6.1.24. (iv) $\mathcal{O}_d = \mathcal{O}_D$.}

10. (i) If $(X_1, d_1), \dots, (X_r, d_r)$ is a finite list of metric spaces (3.1.4) and d is the metric on the product set $X = X_1 \times \dots \times X_r$ defined by

$$d(x, y) = \max_{1 \leq k \leq r} d_k(x_k, y_k)$$

for $x = (x_1, \dots, x_r)$ and $y = (y_1, \dots, y_r)$ (cf. §3.1, Exercise 6), then (X, d) is complete (totally bounded, compact) if and only if all of the spaces $(X_1, d_1), \dots, (X_r, d_r)$ are complete (totally bounded, compact).

(ii) If d and d' are metrics on a set X , and if there exist constants $\alpha > 0, \beta > 0$ such that

$$\alpha d(x, y) \leq d'(x, y) \leq \beta d(x, y)$$

for all $x, y \in X$ (so that, in particular, d and d' are equivalent metrics by Corollary 3.3.6), then (X, d) is complete (totally bounded, compact) if and only if (X, d') is complete (totally bounded, compact).

(iii) Same as Part (i), with d defined by

$$d(x, y) = \left(\sum_{k=1}^r [d_k(x_k, y_k)]^p \right)^{1/p},$$

where $p \geq 1$ is a constant.

{Hint: Cf. Example 3.2.2, (v).}

(iv) Let $(X_k, d_k)_{k \in \mathbb{P}}$ be a sequence of (nonempty) metric spaces and, as in §3.1, Exercise 7, let d be the metric on the product set $X = \prod_{k=1}^{\infty} X_k$ defined by

$$d(x, y) = \sum_{k=1}^{\infty} \frac{1}{2^k} \cdot \frac{d_k(x_k, y_k)}{1 + d_k(x_k, y_k)}$$

for $x = (x_k), y = (y_k)$ in X . Then (X, d) is complete (totally bounded, compact) if and only if every (X_k, d_k) is complete (totally bounded, compact).

11. Every closed subset A of a quasicompact space X is quasicompact.

{Hint: $\{X - A\}$ is an open covering of $X - A$.}

See also §8.1, Exercise 1.

12. A metric space (X, d) is totally bounded if and only if, for every $\epsilon > 0$, X is the union of finitely many sets of diameter $\leq \epsilon$.

6.2. Uniform Convergence, Iterated Limits Theorem

When two limiting operations are applied successively to a function (or to a family of functions), the outcome in general depends on the order in which they are applied. The 'iterated limits theorem' is a situation in which the

order does not matter, provided that one of the limiting operations takes place 'uniformly' in the sense to be discussed in this section.

6.2.1. *Definition.* Let T be a nonempty set, (Y, ρ) a metric space, $f_n : T \rightarrow Y$ ($n = 1, 2, 3, \dots$) a sequence of functions on T with values in Y , and let $f : T \rightarrow Y$ be another such function. We say that:

(i) $f_n \rightarrow f$ **pointwise** on T (or that f is the **pointwise limit** of the sequence f_n) if, for each $t \in T$, $f_n(t) \rightarrow f(t)$ in the metric space Y . This means (cf. 3.2.1) that for each $t \in T$ and for every $\epsilon > 0$ there exists an index $N = N_{t, \epsilon}$ (depending on t and ϵ) such that $\rho(f_n(t), f(t)) \leq \epsilon$ for all $n \geq N$; formally,

$$(\forall t \in T) (\forall \epsilon > 0) \exists N = N_{t, \epsilon} \ni n \geq N \Rightarrow \rho(f_n(t), f(t)) \leq \epsilon.$$

(ii) $f_n \rightarrow f$ **uniformly** on T (or that f is the **uniform limit** of the sequence f_n) if, for every $\epsilon > 0$, there exists an index $N = N_\epsilon$ (depending on ϵ) such that

$$n \geq N \Rightarrow \rho(f_n(t), f(t)) \leq \epsilon \text{ for every } t \in T;$$

formally,

$$(\forall \epsilon > 0) \exists N = N_\epsilon \ni n \geq N \Rightarrow \rho(f_n(t), f(t)) \leq \epsilon \quad (\forall t \in T).$$

In the perspective of (i), for every $\epsilon > 0$ an index N can be found that 'works' at every point $t \in T$.

6.2.2. *Remark.* If $f_n \rightarrow f$ uniformly, it is clear that $f_n \rightarrow f$ pointwise; the converse is false (Exercise 1).

6.2.3. *Example.* The Weierstrass polynomial approximation theorem affirms that if $f : [a, b] \rightarrow \mathbb{R}$ is any continuous real-valued function on a closed interval, then there exists a sequence of polynomial functions $p_n : [a, b] \rightarrow \mathbb{R}$ such that $p_n \rightarrow f$ uniformly on $[a, b]$; M.H. Stone's generalization of Weierstrass' theorem will be proved later in this chapter (§6.9).

6.2.4. *Definition.* With notations as in 6.2.1, the sequence of functions $f_n : T \rightarrow Y$ is said to be **pointwise Cauchy** on T if, for each $t \in T$, the sequence $(f_n(t))$ is Cauchy in (Y, ρ) ; formally,

$$(\forall t \in T) (\forall \epsilon > 0) \exists N = N_{t, \epsilon} \ni m, n \geq N \Rightarrow \rho(f_m(t), f_n(t)) \leq \epsilon,$$

also expressed by saying that, for each $t \in T$, $\rho(f_m(t), f_n(t)) \rightarrow 0$ as $m, n \rightarrow \infty$.

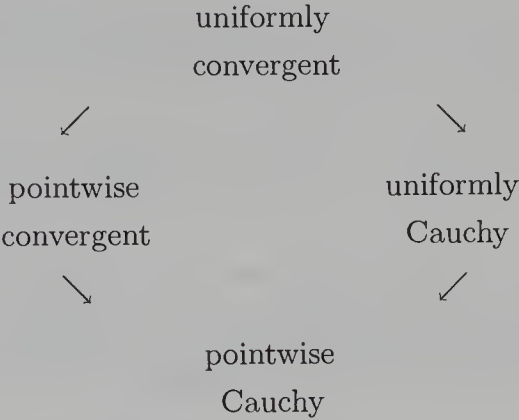
The sequence (f_n) is said to be **uniformly Cauchy** on T if, for every $\epsilon > 0$, there exists an index N such that for $m, n \geq N$, $\rho(f_m(t), f_n(t)) \leq \epsilon$ for all $t \in T$, equivalently, $\sup_{t \in T} \rho(f_m(t), f_n(t)) \leq \epsilon$ whenever $m, n \geq N$; formally,

$$(\forall \epsilon > 0) \exists N = N_\epsilon \ni m, n \geq N \Rightarrow \rho(f_m(t), f_n(t)) \leq \epsilon \quad (\forall t \in T),$$

also expressed by writing

$$\sup_{t \in T} \rho(f_m(t), f_n(t)) \rightarrow 0 \quad \text{as } m, n \rightarrow \infty.$$

6.2.5. *Remark.* A pointwise convergent (uniformly convergent) sequence of functions is pointwise Cauchy (uniformly Cauchy). We thus have the diagram of implications



The following lemma leads up to a ‘Cauchy criterion’ for uniform convergence.

6.2.6. *Lemma.* With notations as in 6.2.1, the following conditions are equivalent:

- (a) (f_n) is uniformly convergent;
- (b) (f_n) is uniformly Cauchy and pointwise convergent.

Proof. (a) \Rightarrow (b): Noted in 6.2.2 and 6.2.5.

(b) \Rightarrow (a): Suppose (f_n) is uniformly Cauchy and $f_n \rightarrow f$ pointwise. By the triangle inequality in Y ,

$$(*) \quad \rho(f_n(t), f(t)) \leq \rho(f_n(t), f_m(t)) + \rho(f_m(t), f(t))$$

for all indices m, n and for all $t \in T$. Given any $\epsilon > 0$, choose an index N such that

$$(i) \quad m, n \geq N \Rightarrow \rho(f_m(t), f_n(t)) \leq \epsilon \text{ for all } t \in T;$$

it will suffice to show that

$$(ii) \quad n \geq N \Rightarrow \rho(f_n(t), f(t)) \leq \epsilon \text{ for all } t \in T.$$

Fix an index $n \geq N$ and fix a point $t \in T$. For all $m \geq N$, it follows from (*) and (i) that

$$\rho(f_n(t), f(t)) \leq \epsilon + \rho(f_m(t), f(t));$$

passing to the limit as $m \rightarrow \infty$, we have $\rho(f_n(t), f(t)) \leq \epsilon + 0$, whence

(ii). \diamond

When the metric space (Y, ρ) of 6.2.1 is complete (6.1.25), every pointwise Cauchy sequence (f_n) is pointwise convergent to a function f , namely the function $f : T \rightarrow Y$ defined by

$$f(t) = \lim_{n \rightarrow \infty} f_n(t) \quad (t \in T).$$

The same is true with “pointwise” replaced by “uniformly”:

6.2.7. *Proposition.* If T is a nonempty set and (Y, ρ) is a complete metric space, then every uniformly Cauchy sequence $f_n : T \rightarrow Y$ ($n = 1, 2, 3, \dots$) is uniformly convergent.

Proof. This is immediate from the preceding remarks and Lemma 6.2.6. \diamond

6.2.8. *Example. (Weierstrass M-test)* Let $\sum_{k=1}^{\infty} M_k$ be a convergent series with terms $M_k \geq 0$ and let (u_k) be a sequence of real-valued (or complex-valued) functions defined on a set T such that, for each index k , $|u_k(t)| \leq M_k$ for all $t \in T$. Then the sequence

$$s_n = \sum_{k=1}^n u_k \quad (n = 1, 2, 3, \dots)$$

is uniformly convergent (and the series $\sum_{k=1}^{\infty} u_k(t)$ is said to be *uniformly convergent* on T). For, if $m < n$ then the inequality

$$|s_m(t) - s_n(t)| = \left| \sum_{k=m+1}^n u_k(t) \right| \leq \sum_{k=m+1}^n M_k,$$

valid for all $t \in T$, shows that the sequence (s_n) is uniformly Cauchy (hence uniformly convergent by the preceding corollary). It follows that the series $\sum |u_k|$ is also uniformly convergent (and one says that the series is ‘uniformly and absolutely convergent’).

As an application, consider a power series $\sum_{k=0}^{\infty} a_k t^k$, where a_0, a_1, a_2, \dots is a sequence of real (or complex) numbers. If c is a real (or complex) number such that the series $\sum_{k=0}^{\infty} a_k c^k$ converges absolutely, then the series $\sum_{k=0}^{\infty} a_k t^k$ is uniformly and absolutely convergent on the closed interval (closed disk) $|t| \leq |c|$. {Here $T = \{t : |t| \leq |c|\}$ and $M_k = |a_k c^k|$.}

If, in the notations of Proposition 6.2.7, the ranges of the functions in question have finite diameter, then the uniform convergence can be expressed in terms of a metric. It is useful to have a short name for such functions:

6.2.9. *Definition.* A function $f : T \rightarrow Y$ with values in a metric space (Y, ρ) is said to be **bounded** if its range $f(T)$ has finite diameter (6.1.14),

that is, if

$$\sup_{s,t \in T} \rho(f(s), f(t)) < +\infty.$$

We write $\mathcal{B} = \mathcal{B}(T, Y)$ for the set of all such functions f (with the understanding that the concept of boundedness depends on the specific metric ρ on Y).

In applications of this concept, it will be useful to know that the union of finitely many sets of finite diameter has finite diameter:

6.2.10. Lemma. *If A and B are sets of finite diameter in the metric space (Y, ρ) , then their union $A \cup B$ also has finite diameter.*

Proof. Let M and N be real numbers ≥ 0 such that

$$\rho(a, a') \leq M \quad \text{and} \quad \rho(b, b') \leq N$$

for all $a, a' \in A$ and all $b, b' \in B$. Fix a pair of points $a' \in A$, $b' \in B$. Given any $x, y \in A \cup B$, we assert that

$$\rho(x, y) \leq M + \rho(a', b') + N;$$

this is obvious if $x, y \in A$ or if $x, y \in B$, whereas if $x \in A$ and $y \in B$ then

$$\rho(x, y) \leq \rho(x, a') + \rho(a', b') + \rho(b', y) \leq M + \rho(a', b') + N. \quad \diamond$$

Here is an important situation in which boundedness comes free of charge:

6.2.11. Proposition. *If T is a quasicompact topological space (6.1.3) and (Y, ρ) is a metric space, then every continuous function $f : T \rightarrow Y$ is bounded.*

Proof. By Theorem 6.1.27, $f(T)$ is a quasicompact subset of Y . If, for every $y \in Y$, U_y is the open ball with radius 1, then $(U_y)_{y \in Y}$ is an open covering of Y , hence of $f(T)$, therefore

$$f(T) \subset U_{y_1} \cup \dots \cup U_{y_r}$$

for a suitable finite list of points y_1, \dots, y_r of Y ; since each U_{y_i} has finite diameter (≤ 2) it follows from the lemma that $f(T)$ has finite diameter. \diamond

The next lemmas will show that the uniform convergence of bounded functions can be expressed in terms of a suitable metric; this leads to the construction of many useful complete metric spaces of functions.

Until further notice, we fix the notations of Definition 6.2.9: T is a nonempty set, (Y, ρ) is a metric space, and $\mathcal{B} = \mathcal{B}(T, Y)$ is the set of all bounded functions $f : T \rightarrow Y$.

6.2.12. Lemma. If $f, g \in \mathcal{B}$ then

$$\sup_{t \in T} \rho(f(t), g(t)) < +\infty.$$

Proof. By assumption, the sets $f(T)$ and $g(T)$ have finite diameter, therefore so does $f(T) \cup g(T)$ (6.2.10); it follows that

$$\sup_{s, t \in T} \rho(f(s), g(t)) < +\infty,$$

and the inequality of the lemma is obtained by specializing to $s = t$. \diamond

6.2.13. Definition. With notations as in the lemma, we define

$$D(f, g) = \sup_{t \in T} \rho(f(t), g(t))$$

(sometimes also denoted $D_\infty(f, g)$). In view of the next lemma, we may call D the **sup-metric** on the set \mathcal{B} :

6.2.14. Lemma. With the foregoing notations, D is a metric on the set \mathcal{B} of bounded functions.

Proof. At any rate, the values of D are nonnegative real numbers, and $D(f, g) = D(g, f)$ by the symmetry of the metric ρ . Similarly, $D(f, f) = 0$ by the corresponding property of ρ . If $f \neq g$ then $f(t) \neq g(t)$ for at least one point t , whence $D(f, g) > 0$. Finally, if $f, g, h \in \mathcal{B}$ then

$$\rho(f(t), h(t)) \leq \rho(f(t), g(t)) + \rho(g(t), h(t)) \leq D(f, g) + D(g, h)$$

for all $t \in T$, whence $D(f, h) \leq D(f, g) + D(g, h)$. Thus, D meets the requirements (i)–(iv) of Definition 3.1.4. \diamond

6.2.15. Remark. If $f_n \in \mathcal{B}$ ($n = 1, 2, 3, \dots$) it is clear from Definitions 6.2.4 and 6.2.13 that

$$(f_n) \text{ is uniformly Cauchy} \Leftrightarrow D(f_m, f_n) \rightarrow 0 \text{ as } m, n \rightarrow \infty,$$

and if $f \in \mathcal{B}$ then

$$f_n \rightarrow f \text{ uniformly on } T \Leftrightarrow D(f_n, f) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Thus, for sequences of bounded functions, ‘uniformly Cauchy’ and ‘uniformly convergent’ are nothing more than the concepts of ‘Cauchy’ and ‘convergent’, respectively, in the metric space (\mathcal{B}, D) .

When the metric space (Y, ρ) is complete, so is the metric space (\mathcal{B}, D) :

6.2.16. Theorem. If T is a nonempty set, (Y, ρ) is a complete metric space, $\mathcal{B} = \mathcal{B}(T, Y)$ is the set of all bounded functions on T with values in Y , and D is the sup-metric on \mathcal{B} (6.2.14), then the metric space (\mathcal{B}, D) is also complete.

Proof. Assuming (f_n) is a sequence in \mathcal{B} with $D(f_m, f_n) \rightarrow 0$ as $m, n \rightarrow \infty$, we seek a function $f \in \mathcal{B}$ such that $D(f_n, f) \rightarrow 0$. Since

(f_n) is uniformly Cauchy (6.2.15), hence pointwise Cauchy, and since Y is complete, we can define a function $f: T \rightarrow Y$ by

$$f(t) = \lim_{n \rightarrow \infty} f_n(t) \quad (t \in T).$$

By definition, $f_n \rightarrow f$ pointwise, therefore $f_n \rightarrow f$ uniformly (6.2.6).

To complete the proof, we need only show that f is bounded. Choose an index N such that

$$\rho(f_N(t), f(t)) \leq 1 \quad \text{for all } t \in T$$

(possible because $f_n \rightarrow f$ uniformly). Since f_N is bounded, there is a positive number K such that

$$\rho(f_N(t), f_N(t')) \leq K \quad \text{for all } t, t' \in T.$$

Then, for all $t, t' \in T$,

$$\begin{aligned} \rho(f(t), f(t')) &\leq \rho(f(t), f_N(t)) + \rho(f_N(t), f_N(t')) + \rho(f_N(t'), f(t')) \\ &\leq 1 + K + 1, \end{aligned}$$

thus $f(T)$ has finite diameter. \diamond

6.2.17. *Example.* If $Y = \mathbb{R}$ with the usual metric, then $\mathcal{B}(T, \mathbb{R})$ is also denoted $\mathcal{B}_{\mathbb{R}}(T)$, and D is the metric

$$D(f, g) = \sup_{t \in T} |f(t) - g(t)| = \|f - g\|_{\infty}$$

considered in Example 3.1.10. Since \mathbb{R} is complete by Cauchy's criterion (1.8.26), by the preceding theorem the set $\mathcal{B}_{\mathbb{R}}(T)$ of all bounded real-valued functions defined on the set T is complete for the sup-metric. The same is true for the set $\mathcal{B}_{\mathbb{C}}(T) = \mathcal{B}(T, \mathbb{C})$ of bounded complex-valued functions.

When the set T is a topological space, we can consider functions $f: T \rightarrow Y$ that are continuous (at particular points, or on all of T). The following theorem shows that the property of continuity is preserved under uniform limits:

6.2.18. *Theorem.* Let T be a topological space, (Y, ρ) a metric space, f and f_n ($n = 1, 2, 3, \dots$) functions on T with values in Y , and let c be a point of T . If (1) each f_n is continuous at c , and (2) $f_n \rightarrow f$ uniformly on T , then f is also continuous at c .

Proof. The following argument recurs so frequently that it might be called the 'classical $\epsilon/3$ proof'. Given any $\epsilon > 0$, we seek a neighborhood V of c such that

$$\rho(f(t), f(c)) \leq \epsilon \quad \text{for all } t \in V$$

(cf. 3.4.3). By the uniformity of the convergence, there exists an index N such that

$$\rho(f_N(t), f(t)) \leq \epsilon/3 \quad \text{for all } t \in T.$$

It follows that for every $t \in T$,

$$\begin{aligned} \rho(f(t), f(c)) &\leq \rho(f(t), f_N(t)) + \rho(f_N(t), f_N(c)) + \rho(f_N(c), f(c)) \\ &\leq \epsilon/3 + \rho(f_N(t), f_N(c)) + \epsilon/3. \end{aligned}$$

By the continuity of f_N at c , there exists a neighborhood V of c such that, for every $t \in V$, $\rho(f_N(t), f_N(c)) \leq \epsilon/3$, whence, by substitution in the preceding inequality,

$$\rho(f(t), f(c)) \leq \epsilon/3 + \epsilon/3 + \epsilon/3 \quad \text{for all } t \in V. \quad \diamond$$

6.2.19. Corollary. Let T be a topological space, (Y, ρ) a metric space, $B = \mathcal{B}(T, Y)$ the set of all bounded functions on T with values in Y , and let c be a point of T . Then, the set

$$\{f \in B : f \text{ is continuous at } c\}$$

is a closed subset of B for the sup-metric.

Proof. This is immediate from the preceding theorem and Remark 6.2.15. \diamond

6.2.20. Definition. If X and Y are topological spaces, we write $\mathcal{C}(X, Y)$ for the set of all continuous functions $f : X \rightarrow Y$. When $Y = \mathbb{R}$ or $Y = \mathbb{C}$ (with the usual absolute-value metric) we also write $\mathcal{C}_{\mathbb{R}}(X) = \mathcal{C}(X, \mathbb{R})$ and $\mathcal{C}_{\mathbb{C}}(X) = \mathcal{C}(X, \mathbb{C})$.

6.2.21. Corollary. Let X be a topological space, (Y, ρ) a metric space, $B = \mathcal{B}(X, Y)$ the set of all bounded functions equipped with the sup-metric D (6.2.14), and $\mathcal{C} = \mathcal{C}(X, Y)$ the set of all continuous functions (6.2.20). Then:

(1) $B \cap \mathcal{C}$ is a closed subset of the metric space B .

(2) If (Y, ρ) is a complete metric space then the set $B \cap \mathcal{C}$ of bounded continuous functions, equipped with the sup-metric, is also a complete metric space.

Proof. (1) The set

$$B \cap \mathcal{C} = \bigcap_{x \in X} \{f \in B : f \text{ is continuous at } x\}$$

is, by the preceding corollary, the intersection of a family of closed sets in B , hence is itself a closed set (3.2.8).

(2) If, moreover, Y is complete, then so is B (6.2.16), hence so is its closed subset $B \cap \mathcal{C}$ (6.1.29). \diamond

6.2.22. Corollary. *With notations as in the preceding corollary, assume in addition that X is quasicompact. Then:*

(1) \mathcal{C} is a closed subset of \mathcal{B} .

(2) If (Y, ρ) is a complete metric space then the set \mathcal{C} , equipped with the sup-metric, is also a complete metric space.

Proof. By the quasicompactness of X , $\mathcal{C} \subset \mathcal{B}$ (6.2.11), so the present corollary simply restates the conclusions of the preceding one. \diamond

6.2.23. Examples. If X is quasicompact then the spaces $\mathcal{C}_{\mathbb{R}}(X)$ and $\mathcal{C}_{\mathbb{C}}(X)$ are complete for the sup-metric $\|f - g\|_{\infty}$. In particular, for every closed interval $[a, b]$ the spaces $\mathcal{C}_{\mathbb{R}}[a, b]$ and $\mathcal{C}_{\mathbb{C}}[a, b]$ are complete for this metric.

The next theorem is important for integration theory in topological spaces¹; it is a (rare!) situation in which pointwise convergence implies uniform convergence:

6.2.24. Theorem. (**Dini's theorem**) *Let X be a quasicompact space, $f \in \mathcal{C} = \mathcal{C}_{\mathbb{R}}(X)$ and $f_n \in \mathcal{C}$ a sequence such that $f_n \rightarrow f$ pointwise on X . If, moreover, $f_1 \leq f_2 \leq f_3 \leq \dots$, then $f_n \rightarrow f$ uniformly on X .*

Proof. By the preceding corollary, \mathcal{C} is complete for the sup-metric $D(g, h) = \|g - h\|_{\infty}$. Let $g_n = f - f_n$; we have $g_n \downarrow 0$ pointwise, and the problem is to show that $g_n \rightarrow 0$ uniformly, that is, $\|g_n\|_{\infty} \rightarrow 0$.

Given any $\epsilon > 0$, we seek an index N such that $\|g_n\|_{\infty} \leq \epsilon$ for all $n \geq N$. For each $x \in X$, $0 \leq g_n(x) < \epsilon$ ultimately, that is, $x \in g_n^{-1}((-\epsilon, \epsilon))$ ultimately. Each of the sets

$$U_n = g_n^{-1}((-\epsilon, \epsilon)) \quad (n = 1, 2, 3, \dots)$$

is open in X by the continuity of g_n , and $U_1 \subset U_2 \subset U_3 \subset \dots$ because $g_1 \geq g_2 \geq g_3 \geq \dots$. By the preceding remark, each $x \in X$ belongs to some (hence to all subsequent) U_n , thus

$$X = \bigcup_{n=1}^{\infty} U_n.$$

Since X is quasicompact,

$$X = U_{n_1} \cup \dots \cup U_{n_r}$$

for a suitable finite set of indices; writing $N = \max\{n_1, \dots, n_r\}$, we have $X = U_N$ (because the sequence U_n is increasing) and $X = U_n$ for all $n \geq N$ (for the same reason). Thus, if $n \geq N$ then $X = g_n^{-1}((-\epsilon, \epsilon))$, in other words $g_n(x) < \epsilon$ for all $x \in X$, whence $\|g_n\|_{\infty} \leq \epsilon$ (in fact, $< \epsilon$). \diamond

¹ Cf. E. Hewitt and K. Stromberg, *Real and abstract analysis* [Springer-Verlag, New York, 1965], p. 115, (9.6) and p. 205, (13.40).

The next application of uniformity gives conditions under which the order of two limiting operations can be reversed; the basic setup is the one already employed in the definition of the limit of a function (3.5.1):

6.2.25. Theorem. (**Iterated limits theorem**) Let (X, d) be a metric space, (Y, ρ) a complete metric space. Suppose

$$(i) \quad A \subset B \subset X, \quad c \in \overline{A},$$

and let

$$(ii) \quad f : B \rightarrow Y \quad \text{and} \quad f_n : B \rightarrow Y \quad (n = 1, 2, 3, \dots)$$

be functions such that $f_n \rightarrow f$ uniformly on A , that is,

$$(iii) \quad f_n|_A \rightarrow f|_A \quad \text{uniformly.}$$

Finally, suppose each f_n has a limit y_n as x approaches c through values in A , that is (cf. 3.5.1),

$$(iv) \quad \exists \lim_{x \rightarrow c, x \in A} f_n(x) = y_n \in Y \quad (n = 1, 2, 3, \dots).$$

Then:

- (1) The sequence (y_n) is convergent in Y , and
- (2) f has a limit equal to $\lim_{n \rightarrow \infty} y_n$ as x approaches c through values in A , that is,

$$\exists \lim_{x \rightarrow c, x \in A} f(x) = \lim_{n \rightarrow \infty} y_n;$$

thus,

$$\lim_{x \rightarrow c, x \in A} \left(\lim_{n \rightarrow \infty} f_n(x) \right) = \lim_{n \rightarrow \infty} \left(\lim_{x \rightarrow c, x \in A} f_n(x) \right).$$

Proof. It is the latter formula that gives the theorem its name. It is helpful to have a picture of the setup underlying (iv):

$$\begin{array}{ccc} & X & \\ & \cup & \\ & B & \xrightarrow{f_n} Y \ni y_n \\ & \cup & \\ c \in \overline{A} \supset A & & \end{array}$$

We consider two cases, according as c does or does not belong to A .

case 1: $c \in A$.

In this case, statement (iv) says that for each n , $y_n = f_n(c)$ and $f_n|_A$ is continuous at c (for the restricted metric $d|_A \times A$; cf. 3.5.2). Then $y_n = f_n(c) \rightarrow f(c)$ and, by Theorem 6.2.18, $f|_A$ is continuous at c , thus

$$\exists \lim_{x \rightarrow c, x \in A} f(x) = f(c) = \lim_{n \rightarrow \infty} y_n.$$

case 2: $c \notin A$.

(1) Since Y is complete, to prove (1) it suffices to show that the sequence (y_n) is Cauchy. Given any $\epsilon > 0$, we seek an index N such that $\rho(y_m, y_n) \leq \epsilon$ for all $m, n \geq N$. By (iii) there exists an index N such that

$$m, n \geq N \Rightarrow \rho(f_m(x), f_n(x)) \leq \epsilon \text{ for all } x \in A$$

(cf. 6.2.6). Fix a pair of indices $m, n \geq N$. For all $x \in A$,

$$\begin{aligned} (*) \quad \rho(y_m, y_n) &\leq \rho(y_m, f_m(x)) + \rho(f_m(x), f_n(x)) + \rho(f_n(x), y_n) \\ &\leq \rho(y_m, f_m(x)) + \epsilon + \rho(f_n(x), y_n). \end{aligned}$$

Since $c \in \bar{A}$ we may choose a sequence (x_k) in A such that $x_k \rightarrow c$. By (iv), $\rho(y_m, f_m(x_k)) \rightarrow 0$ and $\rho(y_n, f_n(x_k)) \rightarrow 0$ as $k \rightarrow \infty$ (3.5.1). Replacing x by x_k in $(*)$ and letting $k \rightarrow \infty$, we have $\rho(y_m, y_n) \leq 0 + \epsilon + 0$, as we wished to show.

(2) Let $y = \lim_{n \rightarrow \infty} y_n$ and, for each n , define $\bar{f}_n : A \cup \{c\} \rightarrow Y$ by the formula

$$\bar{f}_n(x) = \begin{cases} y_n & \text{for } x = c \\ f_n(x) & \text{for } x \in A. \end{cases}$$

If $x_k \in A \cup \{c\}$ and $x_k \rightarrow c$ then $f_n(x_k) \rightarrow y_n = \bar{f}_n(c)$ as $k \rightarrow \infty$, thus \bar{f}_n is continuous at c . Similarly, define $\bar{f} : A \cup \{c\} \rightarrow Y$ by

$$\bar{f}(x) = \begin{cases} y & \text{for } x = c \\ f(x) & \text{for } x \in A. \end{cases}$$

Since $f_n|_A \rightarrow f|_A$ uniformly and $\bar{f}_n(c) = y_n \rightarrow y = \bar{f}(c)$, it is clear that $\bar{f}_n \rightarrow \bar{f}$ uniformly on $A \cup \{c\}$; by case 1,

$$\exists \lim_{x \rightarrow c, x \in A \cup \{c\}} \bar{f}(x) = \bar{f}(c) = y,$$

and, since $\bar{f} = f$ on A we see that

$$\exists \lim_{x \rightarrow c, x \in A} f(x) = y. \diamond$$

6.2.26. Corollary. Let $f_n : [a, b] \rightarrow \mathbb{R}$ ($n = 1, 2, 3, \dots$) and $f : [a, b] \rightarrow \mathbb{R}$ be functions such that $f_n \rightarrow f$ uniformly on $[a, b]$, and let $a \leq c < b$. If every f_n has a right limit at c then so does f , and $f(c+) = \lim_{n \rightarrow \infty} f_n(c+)$.

Proof. Here $X = B = [a, b]$ and $A = (c, b]$. We know that $f_n \rightarrow f$ uniformly on A and, for every n ,

$$\exists \lim_{x \rightarrow c, x \in A} f_n(x) = f_n(c+) \in \mathbb{R},$$

so by the iterated limits theorem, the sequence $(f_n(c+))$ is convergent and

$$\exists \lim_{x \rightarrow c, x \in A} f(x) = \lim_{n \rightarrow \infty} f_n(c+). \quad \diamond$$

6.2.27. Corollary. (**Term-by-term-differentiation**) Let $s_n : [a, b] \rightarrow \mathbb{R}$ be a sequence of functions such that, for every n , s_n is differentiable on $[a, b]$ (one-sided at the endpoints). Assume that there exists a function $t : [a, b] \rightarrow \mathbb{R}$ such that $s'_n \rightarrow t$ uniformly on $[a, b]$; finally, assume that there exists a point $c \in [a, b]$ such that the sequence $(s_n(c))$ is convergent.

Then, the sequence (s_n) is uniformly convergent on $[a, b]$, the limit function $s = \lim s_n$ is differentiable on $[a, b]$ (one-sided at the endpoints) and $s' = t$. Thus,

$$(\lim s_n)' = \lim s'_n.$$

Proof. Note that each s_n is continuous (even differentiable) on $[a, b]$. Convention: we are writing $s'_n(a)$ for the right-derivative $(s_n)'_r(a)$, and $s'_n(b)$ for the left-derivative $(s_n)'_l(b)$. Since the space $\mathcal{C} = \mathcal{C}_{\mathbb{R}}[a, b]$ is complete for the sup-metric (6.2.23), to prove that (s_n) is uniformly convergent we need only show that it is uniformly Cauchy.

Let $\epsilon > 0$. Since the sequence of derivatives (s'_n) is uniformly Cauchy (indeed, uniformly convergent to t) there is an index N such that

$$(i) \quad m, n \geq N \Rightarrow |s'_m(x) - s'_n(x)| \leq \frac{\epsilon}{2(b-a)} \quad \text{for all } x \in [a, b].$$

While we are at it, we can suppose that also

$$(ii) \quad m, n \geq N \Rightarrow |s_m(c) - s_n(c)| \leq \frac{\epsilon}{2}.$$

Fix $m, n \geq N$; it will suffice to show that

$$(*) \quad |s_m(x) - s_n(x)| \leq \epsilon \quad \text{for all } x \in [a, b].$$

Fix $x \in [a, b]$. If $x = c$ the inequality of $(*)$ holds by (ii). Suppose, for example, that $x > c$ (if $x < c$ the argument is similar). By the mean value theorem applied to the function $s_m - s_n$ on the interval $[c, x]$, there exists a point $\xi \in (c, x)$ such that

$$(s_m - s_n)(x) - (s_m - s_n)(c) = (s_m - s_n)'(\xi) \cdot (x - c);$$

it then follows from (i) that

$$(iii) \quad \begin{aligned} |(s_m - s_n)(x) - (s_m - s_n)(c)| &\leq \frac{\epsilon}{2(b-a)} \cdot (x - c) \\ &\leq \frac{\epsilon}{2(b-a)} \cdot (b - a) = \frac{\epsilon}{2}, \end{aligned}$$

thus

$$\begin{aligned} |(s_m - s_n)(x)| &\leq |(s_m - s_n)(x) - (s_m - s_n)(c)| + |(s_m - s_n)(c)| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \end{aligned}$$

which proves (*).

Let $s \in \mathcal{C}$ be the function to which the sequence (s_n) converges uniformly. Note that the hypothesis that $(s_n(c))$ is convergent for at least one point c has now been strengthened to ‘for every point c ’, thus we need only show that $s'(c)$ exists and is equal to $t(c)$. In the framework of Theorem 6.2.25 we contemplate $X = [a, b]$ and $B = A = [a, b] - \{c\}$. Define the ‘difference-quotient’ functions $f_n, f : A \rightarrow \mathbb{R}$ by

$$f_n(x) = \frac{s_n(x) - s_n(c)}{x - c}, \quad f(x) = \frac{s(x) - s(c)}{x - c}.$$

For every n , by assumption

$$\exists \lim_{x \rightarrow c, x \in A} f_n(x) = s'_n(c);$$

we are to show that

$$\exists \lim_{x \rightarrow c, x \in A} f(x) = t(c).$$

By Theorem 6.2.25 it will suffice to show that $f_n \rightarrow f$ uniformly on A .

At any rate, $f_n \rightarrow f$ pointwise on A (because $s_n \rightarrow s$), so it will suffice by Lemma 6.2.6 to show that the sequence (f_n) is uniformly Cauchy. Given any $\epsilon > 0$, choose the index N as earlier in the proof; if $m, n \geq N$, it follows from the first inequality in (iii) (with x no longer required to be $> c$) that

$$|(s_m - s_n)(x) - (s_m - s_n)(c)| \leq \frac{\epsilon}{2(b-a)} \cdot |x - c| \quad \text{for all } x \in [a, b],$$

that is,

$$|[s_m(x) - s_m(c)] - [s_n(x) - s_n(c)]| \leq \frac{\epsilon}{2(b-a)} \cdot |x - c| \quad \text{for all } x \in [a, b];$$

if $x \in A$ then $x \neq c$ and we can divide by $|x - c|$, thus

$$m, n \geq N \Rightarrow |f_m(x) - f_n(x)| \leq \frac{\epsilon}{2(b-a)} \quad \text{for all } x \in A.$$

This shows that (f_n) is uniformly Cauchy on A and completes the proof of the corollary. \diamond

6.2.28. *Example.* As in 6.2.8, let $\sum_{k=0}^{\infty} a_k x^k$ be a power series with real coefficients a_k . Assume, in addition, that the sequence $|a^k|^{1/k}$ is bounded, so that the interval of convergence of the series is nondegenerate (cf. §1.16, Exercise 3). Let $[a, b]$ be a closed interval contained in the interior of the

interval of convergence, and define functions $s_n, s, t : [a, b] \rightarrow \mathbb{R}$ by the formulas

$$s_n(x) = \sum_{k=0}^n a_k x^k, \quad s(x) = \sum_{k=0}^{\infty} a_k x^k, \quad t(x) = \sum_{k=0}^{\infty} (k+1) a_{k+1} x^k$$

(one sees from the limit $k^{1/k} \rightarrow 1$ that the series defining t has the same radius of convergence as the series defining s , of which it is the formal term-by-term derivative). As noted in 6.2.8, $s_n \rightarrow s$ uniformly on $[a, b]$, and, by the same reasoning, $s'_n \rightarrow t$ uniformly on $[a, b]$. It follows from the preceding corollary that s is differentiable on $[a, b]$ with $s' = t$. Given the freedom of choosing $[a, b]$, we see that if R is the radius of convergence of the power series $\sum_{k=0}^{\infty} a_k x^k$, then the function defined by the series is differentiable at every point of the open interval $(-R, R)$, and its derivative may be calculated term by term.

Exercises

1. For each $n = 1, 2, 3, \dots$ let $f_n : \mathbb{Z} \rightarrow \mathbb{R}$ be the function

$$f_n(k) = \begin{cases} 1 & \text{if } k = n \\ 0 & \text{if } k \neq n. \end{cases}$$

Equip \mathbb{R} with its usual metric. Then $f_n \rightarrow 0$ pointwise on \mathbb{Z} , but the convergence is not uniform.

2. If $f_n \rightarrow f$ uniformly and f is continuous at c , it can happen that every f_n is discontinuous at c .

3. (i) If T is a nonempty set and (Y, ρ) is a metric space of finite diameter, then $\mathcal{B}(T, Y)$ coincides with the set $\mathcal{F} = \mathcal{F}(T, Y)$ of all functions $f : T \rightarrow Y$. For sequences of functions $f_n \in \mathcal{F}$, the concepts of 'uniformly convergent' and 'uniformly Cauchy' can be expressed by the sup-metric D as in 6.2.15.

(ii) If, moreover, (Y, ρ) is complete, then \mathcal{F} is complete for the sup-metric D . Application: (Y, ρ) any compact metric space (cf. 6.1.26). Example: $Y = [a, b]$ with the usual metric.

4. A nonempty subset A of a metric space (Y, ρ) is said to be *bounded* if it has finite diameter in the sense of 6.1.14, equivalently, the insertion mapping $i : A \rightarrow Y$ is bounded in the sense of Definition 6.2.9. By convention, the empty subset of Y is also bounded.

Let n be a positive integer, p a real number ≥ 1 , and equip \mathbb{C}^n with the metric derived from the Minkowski p -norm (3.1.14). Prove: A subset A of \mathbb{C}^n is compact if and only if it is closed and bounded. (The compact subsets of \mathbb{R}^n are characterized similarly.)

{Hint: Theorem 6.1.23 and §6.1, Exercise 3.}

6.3. Complete Metric Spaces

Recall that a metric space is said to be *complete* if every Cauchy sequence in the space is convergent (6.1.25). Here is another characterization of completeness:

6.3.1. Theorem. *The following conditions on a metric space (X, d) are equivalent:*

- (a) X is complete;
- (b) X has the **nested closed sets property**: if (F_n) is a sequence of nonempty closed sets such that

$$F_1 \supset F_2 \supset F_3 \supset \dots \quad \text{and} \quad \text{diam } F_n \rightarrow 0,$$

then the F_n have exactly one point in common, that is,

$$\bigcap_{n=1}^{\infty} F_n = \{x\}$$

for some point x .

Proof. Sets of finite diameter were defined in 6.1.14. In general if A is any nonempty subset of X , one defines

$$\text{diam } A = \sup_{x, y \in A} d(x, y)$$

(with $+\infty$ as an admissible value for the supremum); inasmuch as condition (b) requires that $\text{diam } F_n \rightarrow 0$ (in the space \mathbb{R} of 3.3.17) we need only consider subsets of finite diameter.

(a) \Rightarrow (b): Let (F_n) be a decreasing sequence of nonempty closed sets with diameters tending to 0. The desired common point x will be obtained as the limit of a suitable convergent sequence.

For each index n , choose a point $x_n \in F_n$. The sequence (x_n) is Cauchy; for, if $\epsilon > 0$ and N is an index such that $\text{diam } F_N \leq \epsilon$ then, for all $m, n \geq N$, we have $x_m \in F_m \subset F_N$ and $x_n \in F_n \subset F_N$, whence $d(x_m, x_n) \leq \text{diam } F_N \leq \epsilon$. Since X is complete, $x_n \rightarrow x$ for a suitable point x ; we will show that $\bigcap F_n = \{x\}$ by a 'double-inclusion' argument.

For each index n , $k \geq n \Rightarrow x_k \in F_k \subset F_n$; since F_n is closed, we conclude that $x = \lim x_k \in F_n$. Thus $x \in \bigcap F_n$, that is, $\{x\} \subset \bigcap F_n$. On the other hand, if $y \in F_n$ for all n then $d(x, y) \leq \text{diam } F_n \rightarrow 0$, therefore $y = x$; thus $\bigcap F_n \subset \{x\}$.

(b) \Rightarrow (a): Let (x_n) be a Cauchy sequence in X ; the desired limit x for the sequence will be found by applying (b) to a suitable sequence (F_n) of closed sets. For each positive integer n , let

$$F_n = \overline{\{x_k : k \geq n\}}$$

(the set of points that can be approximated as closely as we like by terms x_k of index $k \geq n$); each F_n is a nonempty closed set and $F_1 \supset F_2 \supset F_3 \supset \dots$ (cf. 3.3.16).

We assert that $\text{diam } F_n \rightarrow 0$. At any rate, $\text{diam } F_n \downarrow$, so given any $\epsilon > 0$ it will suffice to find an index N such that $\text{diam } F_N \leq 3\epsilon$. Since (x_n) is Cauchy, we can choose an index N such that

$$m, n \geq N \Rightarrow d(x_m, x_n) \leq \epsilon;$$

given any $x, y \in F_N$ it will suffice to show that $d(x, y) \leq 3\epsilon$. By the definition of F_N , there exist indices $m, n \geq N$ such that

$$d(x, x_m) \leq \epsilon \text{ and } d(y, x_n) \leq \epsilon;$$

then $d(x_m, x_n) \leq \epsilon$ by the choice of N , therefore

$$d(x, y) \leq d(x, x_m) + d(x_m, x_n) + d(x_n, y) \leq \epsilon + \epsilon + \epsilon.$$

Thus, the sets F_n satisfy the conditions of (b). Let x be their common point. For every index n , both x_n and x belong to F_n , therefore $d(x_n, x) \leq \text{diam } F_n$, whence $d(x_n, x) \rightarrow 0$. \diamond

The concept of continuity in a topological space is expressible entirely in terms of open sets, without reference to any metric from which the topology might have been derived (cf. 3.4.2, 3.4.5). In a metric space, there is a stronger concept that does make reference to the metric:

6.3.2. Definition. Let (X, d) and (Y, ρ) be metric spaces. A function $f : X \rightarrow Y$ is said to be **uniformly continuous** on X if, for every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$d(x, x') < \delta \Rightarrow \rho(f(x), f(x')) < \epsilon,$$

that is, any two points in X that are within δ of each other have images in Y that are within ϵ of each other. Formally,

$$(\forall \epsilon > 0) \exists \delta > 0 \ni x, x' \in X, d(x, x') < \delta \Rightarrow \rho(f(x), f(x')) < \epsilon.$$

6.3.3. Remark. Every uniformly continuous function is continuous. For, continuity at each point x' means that for every $\epsilon > 0$ there exists a $\delta > 0$, depending in general on both ϵ and x' , for which

$$d(x, x') < \delta \Rightarrow \rho(f(x), f(x')) < \epsilon,$$

whereas uniform continuity on X means that for every $\epsilon > 0$, a $\delta > 0$ can be found that 'works' at every point x' of X .

6.3.4. Proposition. Let (X, d) and (Y, ρ) be metric spaces, $f : X \rightarrow Y$. Consider the following conditions:

- (a) f is uniformly continuous on X ;
- (b) for sequences $(x_n), (x'_n)$ in X ,

$$d(x_n, x'_n) \rightarrow 0 \Rightarrow \rho(f(x_n), f(x'_n)) \rightarrow 0;$$

(c) for a sequence (x_n) in X ,

$$(x_n) \text{ Cauchy in } X \Rightarrow (f(x_n)) \text{ Cauchy in } Y.$$

Then $(a) \Leftrightarrow (b) \Rightarrow (c)$.

Proof. $(a) \Rightarrow (c)$: Let (x_n) be a Cauchy sequence in X . Given any $\epsilon > 0$, we are to show that $\rho(f(x_m), f(x_n)) < \epsilon$ for all sufficiently large m and n . Choose $\delta > 0$ as in Definition 6.3.2, then choose an index N such that $d(x_m, x_n) < \delta$ for all $m, n \geq N$; it then follows from the choice of δ that $\rho(f(x_m), f(x_n)) < \epsilon$ for all $m, n \geq N$. Incidentally, the reverse implication fails (Exercise 5).

$(a) \Rightarrow (b)$: Let (x_n) and (x'_n) be sequences in X such that $d(x_n, x'_n) \rightarrow 0$. Given any $\epsilon > 0$, we are to show that $\rho(f(x_n), f(x'_n)) < \epsilon$ ultimately. Choose $\delta > 0$ as in 6.3.2, then choose an index N such that $d(x_n, x'_n) < \delta$ for all $n \geq N$; by the choice of δ , $\rho(f(x_n), f(x'_n)) < \epsilon$ for all $n \geq N$.

$(b) \Rightarrow (a)$: We argue contrapositively: assuming f is not uniformly continuous, let us construct a pair of sequences $(x_n), (x'_n)$ in X for which the implication in (b) fails. By assumption, there exists an $\epsilon > 0$ such that, for every $\delta > 0$, the implication of Definition 6.3.2 fails; in particular, for each positive integer n , the implication fails for $\delta = 1/n$, thus there exists a pair of points x_n, x'_n in X such that $d(x_n, x'_n) < 1/n$ but $\rho(f(x_n), f(x'_n)) \geq \epsilon$. Then $d(x_n, x'_n) \rightarrow 0$ but $\rho(f(x_n), f(x'_n)) \not\rightarrow 0$. \diamond

The function $f : (0, 1] \rightarrow \mathbb{R}$ defined by $f(x) = \sin(1/x)$ is continuous on its domain, but the right limit $f(0+)$ fails to exist (for example, there are sequences $x_n \rightarrow 0$ and $y_n \rightarrow 0$ in $(0, 1]$ with $f(x_n) = 0$ and $f(y_n) = 1$ for all n); thus, no definition of $f(0)$ can render f continuous on the closed interval $[0, 1]$. It is easy to see directly that f fails to be uniformly continuous, but the following theorem—a capital application of uniform continuity—gives an interesting roundabout way of seeing it.

6.3.5. Theorem. *If $f : A \rightarrow Y$ is a **uniformly continuous** function defined on a dense subset A of a metric space (X, d) and taking values in a **complete** metric space (Y, ρ) , then there exists a unique continuous function $\bar{f} : X \rightarrow Y$ such that $\bar{f}|_A = f$; moreover, the function \bar{f} is uniformly continuous.*

Proof. Uniqueness. Assuming g and h are continuous functions from X into Y whose restrictions to A are equal to f (hence to each other), we are to show that $g = h$ on X . Writing

$$B = \{x \in X : g(x) = h(x)\},$$

we know that $A \subset B$ and that $g = h$ on B ; we are to show that $B = X$. The set B is closed in X ; for, if $x_n \in B$ and $x_n \rightarrow x \in X$ then, by the continuity of g and h ,

$$g(x) = \lim g(x_n) = \lim h(x_n) = h(x),$$

so that $x \in B$. It follows that $X = \bar{A} \subset B$, whence $g = h$ on X .

Existence. The function $\bar{f} : X \rightarrow Y$ will be defined by specifying its graph G (cf. 1.3.1). Let

$$G = \{(x, y) \in X \times Y : \exists a_n \in A \text{ with } a_n \rightarrow x \text{ and } f(a_n) \rightarrow y\};$$

to show that G is the graph of a function, we must show that for each $x \in X$ there exists one and only one $y \in Y$ such that $(x, y) \in G$. Given any $x \in X$, choose a sequence (a_n) in A with $a_n \rightarrow x$ (possible by the density of A); since (a_n) is Cauchy and f is uniformly continuous, the sequence $(f(a_n))$ is also Cauchy (6.3.4) and therefore convergent in Y , thus $(x, y) \in G$ with $y = \lim f(a_n)$. Note that if (a'_n) is any other sequence in A with $a'_n \rightarrow x$, then necessarily $f(a'_n) \rightarrow y$; for,

$$d(a'_n, a_n) \leq d(a'_n, x) + d(x, a_n) \rightarrow 0 + 0,$$

therefore $\rho(f(a'_n), f(a_n)) \rightarrow 0$ (6.3.4), thus

$$\rho(f(a'_n), y) \leq \rho(f(a'_n), f(a_n)) + \rho(f(a_n), y) \rightarrow 0 + 0.$$

It follows that y is the *only* point of Y for which $(x, y) \in G$. We may therefore define a function $\bar{f} : X \rightarrow Y$ by

$$\bar{f}(x) = y, \text{ where } (x, y) \in G,$$

and the above argument shows that if (a_n) is a sequence in A with $a_n \rightarrow x \in X$ then $\bar{f}(x) = \lim f(a_n)$. In particular, if $a \in A$ and (a_n) is the constant sequence $a_n = a$, then

$$\bar{f}(a) = \lim f(a_n) = f(a),$$

thus \bar{f} is an extension of f .

To complete the proof, we need only show that \bar{f} is uniformly continuous on X . Given any $\epsilon > 0$, choose $\delta > 0$ so that

$$a, a' \in A, d(a, a') < \delta \Rightarrow \rho(f(a), f(a')) < \epsilon.$$

Assuming $x, x' \in X$ with $d(x, x') < \delta$, it will suffice to show that $\rho(\bar{f}(x), \bar{f}(x')) \leq \epsilon$. Let (a_n) and (a'_n) be sequences in A such that $a_n \rightarrow x$ and $a'_n \rightarrow x'$. From $d(x, x') < \delta$ and the inequality

$$d(a_n, a'_n) \leq d(a_n, x) + d(x, x') + d(x', a'_n)$$

we see that $d(a_n, a'_n) < \delta$ ultimately, therefore $\rho(f(a_n), f(a'_n)) < \epsilon$ ultimately (by the choice of δ). Thus, for all sufficiently large n ,

$$\begin{aligned} \rho(\bar{f}(x), \bar{f}(x')) &\leq \rho(\bar{f}(x), f(a_n)) + \rho(f(a_n), f(a'_n)) + \rho(f(a'_n), \bar{f}(x')) \\ &< \rho(\bar{f}(x), f(a_n)) + \epsilon + \rho(f(a'_n), \bar{f}(x')); \end{aligned}$$

since $f(a_n) \rightarrow \bar{f}(x)$ and $f(a'_n) \rightarrow \bar{f}(x')$, passage to the limit in the preceding inequality yields $\rho(\bar{f}(x), \bar{f}(x')) \leq 0 + \epsilon + 0$. \diamond

6.3.6. Remark. Informally, *uniform convergence* prolongs continuity to the limit function (6.2.18), while *uniform continuity* prolongs continuity to

the 'limit domain' (6.3.5), in a vague sense justifying the use of the term 'uniform' in the two contexts; but the true justification comes in the theory of 'uniform structures'.¹

The uniformity of continuity is assured when the domain is compact:

6.3.7. Theorem. If $f : X \rightarrow Y$ is a continuous function defined on a compact metric space (X, d) and taking values in a metric space (Y, ρ) , then f is uniformly continuous.

Proof. Assume to the contrary that the continuous function f is not uniformly continuous. By Proposition 6.3.4, there exists a pair of sequences $(x_n), (x'_n)$ in X such that $d(x_n, x'_n) \rightarrow 0$ but $\rho(f(x_n), f(x'_n)) \not\rightarrow 0$. Write $y_n = f(x_n)$, $y'_n = f(x'_n)$. Since $\rho(y_n, y'_n) \not\rightarrow 0$, there exists an $\epsilon > 0$ such that $\rho(y_n, y'_n)$ fails to be $< \epsilon$ ultimately, in other words, such that $\rho(y_n, y'_n) \geq \epsilon$ frequently (the terminology is explained in the proof of 1.16.4). Passing to a subsequence of $(x_n, x'_n) \in X \times X$, we can suppose that $\rho(y_n, y'_n) \geq \epsilon$ for all n . Since X is compact, (x_n) has a convergent subsequence, say $x_{n_k} \rightarrow x$; together with $d(x_{n_k}, x'_{n_k}) \rightarrow 0$, this implies that also $x'_{n_k} \rightarrow x$. Since f is continuous, $f(x_{n_k}) \rightarrow f(x)$ and $f(x'_{n_k}) \rightarrow f(x)$, therefore (cf. 3.2.4)

$$\rho(y_{n_k}, y'_{n_k}) = \rho(f(x_{n_k}), f(x'_{n_k})) \rightarrow \rho(f(x), f(x)) = 0,$$

contrary to $\rho(y_n, y'_n) \geq \epsilon$ for all n . \diamond

In particular, every continuous real-valued function $f : [a, b] \rightarrow \mathbb{R}$ defined on a closed interval is uniformly continuous (cf. 6.1.9), a fact that is often cited in proving the Riemann-integrability of f .

The next circle of ideas deals with a property of the *topology* derived from a complete metric (3.3.2); the following terminology is due to Bourbaki², the classical terminology being indicated parenthetically:

6.3.8. Definition. Let (X, \mathcal{O}) be a topological space (3.3.1), A a subset of X .

We say that A is **rare** (or that 'A is nowhere dense in X ') if its closure \overline{A} has no interior points, equivalently,

$$U \subset \overline{A}, U \in \mathcal{O} \Rightarrow U = \emptyset.$$

We say that A is **meager** (or that 'A is of the first category in X ') if it is the union of a sequence of rare subsets of X , that is,

$$A = \bigcup_{n=1}^{\infty} A_n, \quad \text{int } \overline{A_n} = \emptyset \quad (n = 1, 2, 3, \dots).$$

¹ Cf. N. Bourbaki, *General topology*, Vol. I, Chapter II [Addison-Wesley, Reading, 1966].

² N. Bourbaki, *op. cit.*, Vol. II, Chapter IX.

If A is not a meager subset of X we say that A is **nonmeager** (or that ' A is of the second category in X '); stated contrapositively,

$$A = \bigcup_{n=1}^{\infty} A_n \Rightarrow \text{int } \overline{A_n} \neq \emptyset \text{ for at least one value of } n.$$

The topological space X is said to be a **Baire space** if \emptyset is the only meager open set, that is,

$$U \in \mathcal{O}, U \neq \emptyset \Rightarrow U \text{ nonmeager.}$$

(In particular, every nonempty Baire space is a nonmeager subset of itself—in the classical language, every nonempty Baire space is of the second category.)

6.3.9. Theorem. (**Baire category theorem**) *Let (X, d) be a complete metric space, equipped with the topology derived from the metric d (3.3.2). Then:*

(1) *If A is a meager subset of X , then its complement is dense; that is,*

$$A \text{ meager} \Rightarrow \overline{X - A} = X.$$

(2) *X is a Baire space.*

Proof. (1) By assumption, $A = \bigcup_{n=1}^{\infty} A_n$ with A_n rare for every n . Since $\overline{A_n}$ is also rare (because $\overline{\overline{A_n}} = \overline{A_n}$ has empty interior), the set $B = \bigcup_{n=1}^{\infty} \overline{A_n}$ is also meager. Since $X - B \subset X - A$, it will suffice to show that $X - B$ is dense.

Changing notations, we can suppose that $A = \bigcup_{n=1}^{\infty} A_n$, where the A_n are closed sets with empty interior. We are to show that every point of X is adherent to $X - A$; given any nonempty open set V , it will suffice to show that $V \cap (X - A) \neq \emptyset$. For every n , $V \cap (X - A_n) \neq \emptyset$; for, $V \cap (X - A_n) = \emptyset$ would imply that $V \subset A_n$, contrary to the assumption that A_n has empty interior. Thus

$$\begin{aligned} V \cap (X - A) &= V \cap \left(X - \bigcup_{n=1}^{\infty} A_n \right) \\ &= V \cap \bigcap_{n=1}^{\infty} (X - A_n) = \bigcap_{n=1}^{\infty} V \cap (X - A_n), \end{aligned}$$

where every term of the intersection on the right side is nonempty. A point of $V \cap (X - A)$ will be obtained by applying the nested closed sets property (6.3.1) to a suitable sequence of closed sets in the complete metric space (X, d) .

Let U_1 be an open set such that

$$\emptyset \neq U_1 \subset \overline{U_1} \subset V \text{ and } \text{diam } \overline{U_1} \leq 1.$$

{For example, let $x \in V$ and choose $r > 0$ small enough so that the closed ball $B_r(x)$ with radius r and center x is contained in V , and such that $r < 1/2$; then the open ball $U_1 = U_r(x)$ has diameter ≤ 1 and $\overline{U}_1 \subset B_r(x) \subset V$.}

Since A_1 has empty interior, it cannot contain U_1 , thus the open set $U_1 \cap (X - A_1)$ is nonempty. Arguing as above, there exists an open set U_2 such that

$$\emptyset \neq U_2 \subset \overline{U}_2 \subset U_1 \cap (X - A_1) \quad \text{and} \quad \text{diam } \overline{U}_2 \leq 1/2.$$

Again, A_2 cannot contain U_2 , so that $U_2 \cap (X - A_2)$ is a nonempty open set; let U_3 be an open set such that

$$\emptyset \neq U_3 \subset \overline{U}_3 \subset U_2 \cap (X - A_2) \quad \text{and} \quad \text{diam } \overline{U}_3 \leq 1/3.$$

Continuing recursively, we construct a sequence of nonempty open sets U_1, U_2, U_3, \dots , where U_{n+1} is chosen so that

$$\emptyset \neq U_{n+1} \subset \overline{U}_{n+1} \subset U_n \cap (X - A_n) \quad \text{and} \quad \text{diam } \overline{U}_{n+1} \leq 1/(n+1).$$

In particular $\overline{U}_{n+1} \subset U_n \subset \overline{U}_n$; thus, the sequence of closed sets $F_n = \overline{U}_n$ satisfies the conditions in (b) of Theorem 6.3.1, consequently

$$\bigcap_{n=1}^{\infty} \overline{U}_n = \{x\}$$

for a suitable point x . Then $x \in \overline{U}_1 \subset V$ and, for every positive integer n ,

$$x \in \overline{U}_{n+1} \subset U_n \cap (X - A_n) \subset X - A_n,$$

thus $x \in V \cap \bigcap_{n=1}^{\infty} (X - A_n) = V \cap (X - A)$ and the proof of (1) is complete.

(2) Given a meager open set U in X , we are to show that U is empty. By (1), $\overline{X - U} = X$; but $X - U$ is closed, whence $X - U = X$, $U = \emptyset$. \diamond

This theorem has a corollary with many important applications in functional analysis:

6.3.10. Corollary. (Uniform boundedness principle) *Let (X, d) be a complete metric space and let $\mathcal{E} \subset \mathcal{C}(X, \mathbb{R})$, that is, \mathcal{E} is a set of continuous real-valued functions defined on X . Assume that \mathcal{E} is pointwise bounded on X , in the sense that for each point $x \in X$, the set of values*

$$\mathcal{E}(x) = \{f(x) : f \in \mathcal{E}\}$$

is a bounded set of real numbers. Then \mathcal{E} is uniformly bounded on some nonempty open set, that is, there exists a nonempty open set U in X such that the set of restrictions

$$\mathcal{E}|U = \{f|U : f \in \mathcal{E}\}$$

is uniformly bounded (i.e., bounded for the sup-norm).

Proof. By assumption, for each point $x \in X$ there exists a real number $M_x > 0$ such that

$$|f(x)| \leq M_x \quad \text{for all } f \in \mathcal{E};$$

we seek a nonempty open set U in X and a real number $M > 0$ such that

$$|f(x)| \leq M \quad \text{for all } x \in U \text{ and for all } f \in \mathcal{E},$$

that is, in the notations of 3.1.10,

$$\|f|_U\|_\infty \leq M \quad \text{for all } f \in \mathcal{E}.$$

For each positive integer n , let

$$\begin{aligned} A_n &= \{x \in X : |f(x)| \leq n \text{ for all } f \in \mathcal{E}\} \\ &= \bigcap_{f \in \mathcal{E}} \{x \in X : |f(x)| \leq n\} \\ &= \bigcap_{f \in \mathcal{E}} f^{-1}([-n, n]); \end{aligned}$$

from the continuity of the functions $f \in \mathcal{E}$, we see that A_n is the intersection of a family of closed sets (3.4.5) and is therefore closed. Moreover,

$$X = \bigcup_{n=1}^{\infty} A_n;$$

for, if $x \in X$ and m is a positive integer such that $m \geq M_x$, then

$$|f(x)| \leq M_x \leq m \quad \text{for all } f \in \mathcal{E},$$

whence $x \in A_m$. Since $X = \bigcup_{n=1}^{\infty} A_n$ is a Baire space (6.3.9), hence is nonmeager, there exists an index \bar{M} such that $A_{\bar{M}}$ is not rare; if U is a nonempty open set such that $U \subset \bar{A}_{\bar{M}} = A_{\bar{M}}$ (for example, $U = \text{int } A_{\bar{M}}$), then

$$|f(x)| \leq M \quad \text{for all } x \in U \text{ and } f \in \mathcal{E},$$

thus the functions $f \in \mathcal{E}$ are uniformly bounded on U . \diamond

Completeness is clearly a useful property for a metric space to have; it is a consolation prize to incomplete spaces that every metric space can be regarded as a dense subspace of a complete metric space:

6.3.11. Theorem. (**Completion**) If (X, d) is any metric space, then:

(**Existence**) There exist a complete metric space (\hat{X}, \hat{d}) and a mapping $f : X \rightarrow \hat{X}$ such that

1° f is distance-preserving: $\hat{d}(f(x), f(y)) = d(x, y)$ for all x, y in X , and

2° f has dense range: $\overline{f(X)} = \hat{X}$.

(Uniqueness) If also (Z, ρ) is a complete metric space and $g : X \rightarrow Z$ is a distance-preserving mapping with dense range, then (Z, ρ) may be identified with (\hat{X}, \hat{d}) , in the sense that there exists a distance-preserving bijection $h : \hat{X} \rightarrow Z$ that carries $f(X)$ into $g(X)$, more precisely, $h(f(x)) = g(x)$ for all $x \in X$.

In order that (X, d) be complete, it is necessary and sufficient that f be surjective.

Proof. A distance-preserving mapping is also said to be **isometric** (or to be an **isometry**).

Existence. The construction of \hat{X} from X imitates Cantor's method for constructing the reals from the rationals (§1.8). Let \mathcal{C} be the set of all Cauchy sequences $s = (x_n)$ in X ; for $s = (x_n)$ and $t = (y_n)$ in \mathcal{C} , we write $s = t$ in case $x_n = y_n$ for all n . An equivalence relation \sim in \mathcal{C} is defined by writing $s \sim t$ in case $d(x_n, y_n) \rightarrow 0$ (the transitivity of \sim follows from the triangle inequality for d). We write $\bar{s} = \{r \in \mathcal{C} : r \sim s\}$ for the equivalence class of s in \mathcal{C} , and \hat{X} for the set of all equivalence classes:

$$\hat{X} = \mathcal{C}/\sim = \{\bar{s} : s \in \mathcal{C}\};$$

thus $s \mapsto \bar{s}$ is the quotient mapping $\mathcal{C} \rightarrow \hat{X}$.

Given $u, v \in \hat{X}$ we are to define a distance $\hat{d}(u, v)$. Say $u = \bar{s}$, $v = \bar{t}$, where $s = (x_n)$ and $t = (y_n)$ are Cauchy sequences in X . The sequence $(d(x_n, y_n))$ is Cauchy (therefore convergent) in \mathbb{R} , as we see from the inequality (cf. 3.2.3)

$$|d(x_m, y_m) - d(x_n, y_n)| \leq d(x_m, x_n) + d(y_m, y_n).$$

To justify defining

$$(*) \quad \hat{d}(u, v) = \lim d(x_n, y_n)$$

we must check that the limit is independent of the particular representatives s, t of the equivalence classes u, v : if also $u = \bar{s'}$, $v = \bar{t'}$, where $s' = (x'_n)$ and $t' = (y'_n)$, then the equality

$$\lim d(x_n, y_n) = \lim d(x'_n, y'_n)$$

follows from the inequality

$$|d(x_n, y_n) - d(x'_n, y'_n)| \leq d(x_n, x'_n) + d(y_n, y'_n)$$

and the fact that $d(x_n, x'_n) \rightarrow 0$, $d(y_n, y'_n) \rightarrow 0$.

The function $\hat{d} : \hat{X} \times \hat{X} \rightarrow \mathbb{R}$ defined by the formula $(*)$ is a metric on \hat{X} . For, if $u = \overline{(x_n)}$, $v = \overline{(y_n)}$, $w = \overline{(z_n)}$ then, passing to the limit in the relations

$$d(x_n, x_n) = 0, \quad d(x_n, y_n) \geq 0, \quad d(x_n, y_n) = d(y_n, x_n),$$

$$d(x_n, z_n) \leq d(x_n, y_n) + d(y_n, z_n),$$

we see that

$$\hat{d}(u, u) = 0, \quad \hat{d}(u, v) \geq 0, \quad \hat{d}(u, v) = \hat{d}(v, u),$$

$$\hat{d}(u, w) \leq \hat{d}(u, v) + \hat{d}(v, w),$$

and if $\hat{d}(u, v) = 0$ then $d(x_n, y_n) \rightarrow 0$, whence $(x_n) \sim (y_n)$, $u = v$.

If $x \in X$ we write (x) for the constant sequence x, x, x, \dots ; the mapping $f : X \rightarrow \hat{X}$ defined by

$$f(x) = \overline{(x)} \quad (x \in X)$$

is isometric since, for $x, y \in X$,

$$\hat{d}(f(x), f(y)) = \lim d(x, y) = d(x, y).$$

In particular, f is injective; for, if $f(x) = f(y)$ then $d(x, y) = \hat{d}(f(x), f(y)) = 0$, whence $x = y$.

To prove that $f(X)$ is dense in \hat{X} , we need only observe that if $u = \overline{(x_n)} \in \hat{X}$ then $f(x_n) \rightarrow u$; indeed, if $\epsilon > 0$ and N is chosen so that $d(x_m, x_n) \leq \epsilon$ for all $m, n \geq N$, then

$$m \geq N \Rightarrow \hat{d}(f(x_m), u) = \lim_{k \rightarrow \infty} d(x_m, x_k) \leq \epsilon.$$

The 'existence' proof will be concluded by showing that the metric space (\hat{X}, \hat{d}) is complete. Let (u_n) be a Cauchy sequence in \hat{X} ; we seek a point $u \in \hat{X}$ with $u_n \rightarrow u$. Consider first the special case that $u_n \in f(X)$ for all n . Then $u_n = f(x_n)$ for a unique $x_n \in X$. Since f is isometric, the sequence (x_n) is also Cauchy; writing $u = \overline{(x_n)}$, we have $u_n \rightarrow u$ by the argument in the preceding paragraph. In the general case, for each index n we may choose a point $v_n \in f(X)$ with $\hat{d}(v_n, u_n) < 1/n$. The sequence (v_n) is also Cauchy, since

$$\begin{aligned} \hat{d}(v_m, v_n) &\leq \hat{d}(v_m, u_m) + \hat{d}(u_m, u_n) + \hat{d}(u_n, v_n) \\ &< 1/m + \hat{d}(u_m, u_n) + 1/n \rightarrow 0 \quad \text{as } m, n \rightarrow \infty. \end{aligned}$$

By the special case considered first, (v_n) has a limit u in \hat{X} , and $u_n \rightarrow u$ follows from the computation

$$\hat{d}(u_n, u) \leq \hat{d}(u_n, v_n) + \hat{d}(v_n, u) < 1/n + \hat{d}(v_n, u) \rightarrow 0.$$

Uniqueness. Suppose (Z, ρ) and $g : X \rightarrow Z$ have the properties indicated in the statement of the theorem; we then have a diagram

$$\begin{array}{ccc} X & \xrightarrow{\text{id}} & X \\ f \downarrow & & \downarrow g \\ \hat{X} & \xrightarrow{\quad h \quad} & Z \end{array}$$

for which we seek an isometric bijection $h: \hat{X} \rightarrow Z$ such that $h(f(x)) = g(x)$ for all $x \in X$. {The completed diagram will then be 'commutative' in the sense that the two ways of getting from X to Z will coincide: $h \circ f = g = g \circ \text{id}$.} We begin by defining a mapping $h_0: f(X) \rightarrow Z$ on the dense subset $f(X)$ of \hat{X} , by means of the formula

$$h_0(f(x)) = g(x) \quad (x \in X);$$

since x is uniquely determined by $f(x)$, there is no ambiguity in the definition of h_0 (one says that h_0 is *well-defined*). The mapping h_0 is isometric since, for all $x, y \in X$,

$$\rho(h_0(f(x)), h_0(f(y))) = \rho(g(x), g(y)) = d(x, y) = \hat{d}(f(x), f(y)).$$

In particular, h_0 is uniformly continuous; since it takes values in a complete metric space, it is extendible to a uniformly continuous mapping $h: \hat{X} \rightarrow Z$ (6.3.5). In fact, h is isometric; for, if $u, v \in \hat{X}$ and $(x_n), (y_n)$ are sequences in X such that $f(x_n) \rightarrow u$ and $f(y_n) \rightarrow v$, then $h(f(x_n)) \rightarrow h(u)$ and $h(f(y_n)) \rightarrow h(v)$ by the continuity of h , thus

$$\begin{aligned} \hat{d}(u, v) &= \lim \hat{d}(f(x_n), f(y_n)) \\ &= \lim d(x_n, y_n) \\ &= \lim \rho(g(x_n), g(y_n)) \\ &= \lim \rho(h_0(f(x_n)), h_0(f(y_n))) \\ &= \lim \rho(h(f(x_n)), h(f(y_n))) \\ &= \rho(h(u), h(v)). \end{aligned}$$

By assumption, $h(f(X)) = h_0(f(X)) = g(X)$ is a dense subset of Z , so all the more $h(\hat{X})$ is dense in Z . However, \hat{X} is complete and h is isometric, so $h(\hat{X})$ is a complete metric subspace of Z and is therefore a closed set in Z (6.1.29); $h(\hat{X}) = h(\hat{X}) = Z$, so h is surjective. {Note, incidentally, that h is uniquely determined by the property $h \circ f = g$, since h is continuous and $f(X)$ is dense in \hat{X} .}

Finally, if X is already complete then $f(X) = \hat{X}$ by the argument used in proving h surjective. Conversely, if f is surjective then it is an isometric bijection, so the completeness of \hat{X} implies that of X . \diamond

6.3.12. *Definition.* With notations as in the preceding theorem, the complete metric space (\hat{X}, \hat{d}) is called the **completion** of the metric space (X, d) , and the isometric mapping $f: X \rightarrow \hat{X}$ is called the **embedding** of X into \hat{X} .

6.3.13. *Remarks.* Two metric spaces that are connected by an isometric bijection—such as (in the above notations) the set X equipped with the metric d and the set $f(X)$ equipped with the restricted metric

$\hat{d}|f(X) \times f(X)$ —are essentially ‘equal’ as metric spaces. They are candidates for ‘identification’: throw away (X, d) , retain $(f(X), \hat{d}|f(X) \times f(X))$, and rename $f(X)$ and $\hat{d}|f(X) \times f(X)$ to be X and d . The net effect is that (X, d) can be regarded as a dense subset of a complete metric (\hat{X}, \hat{d}) such that the metric d is obtained by restricting \hat{d} to pairs of points of X ; after this identification, the embedding mapping of the preceding definition becomes the insertion mapping $i: X \rightarrow \hat{X}$, $i(x) = x$ ($\forall x \in X$).

Here are a few applications of the concept of completion:

6.3.14. Proposition. *A metric space is totally bounded if and only if its completion is compact.*

Proof. By the preceding remarks, we may regard the given metric space (X, d) as a dense subset of a complete metric space (Z, ρ) , with $d = \rho|X \times X$.

If Z is compact then it is totally bounded (6.1.12), therefore every subspace of Z —in particular X —is totally bounded. (Cf. 6.1.24 for a proof by overkill. Alternatively, given any $\epsilon > 0$, write $Z = A_1 \cup \dots \cup A_n$ with $\text{diam } A_i \leq \epsilon/2$ for all i , then consider the sets $A_i \cap X$.)

Conversely, if X is totally bounded then its closure Z is also totally bounded: given any $\epsilon > 0$, choose an $\epsilon/2$ -net F for X ; each point of Z is within $\epsilon/2$ of some point of X , hence within ϵ of some point of F . Thus Z is complete and totally bounded, therefore compact (6.1.26). \diamond

6.3.15. Proposition. *Let (X, d) be a complete metric space and let A be a nonempty subset of X , regarded as a metric subspace of (X, d) (6.1.28). Then the completion of the metric space A may be identified with its closure in X (concisely, $\hat{A} = \overline{A}$).*

Proof. Let us write d_A and $d_{\overline{A}}$ for the restrictions of d to $A \times A$ and $\overline{A} \times \overline{A}$, respectively. The metric space $(\overline{A}, d_{\overline{A}})$ is complete (6.1.29); moreover, A is a dense subset of \overline{A} , so by the “uniqueness” part of Theorem 6.3.11, the insertion mapping $f: A \rightarrow \overline{A}$, $f(a) = a$ ($\forall a \in A$), is extendible to an isometric bijection $\hat{A} \rightarrow \overline{A}$. \diamond

6.3.16. Corollary. *A subset A of a complete metric space X is a totally bounded metric subspace of X if and only if its closure \overline{A} is compact.*

Proof. We have the equivalences

$$A \text{ totally bounded} \Leftrightarrow \hat{A} \text{ compact} \Leftrightarrow \overline{A} \text{ compact}$$

by Propositions 6.3.14 and 6.3.15, respectively. \diamond

The rest of the chapter is devoted to some important examples of complete metric spaces.

Exercises

1. Let $X = (0, 1]$ with the usual metric and let $F_n = (0, 1/n]$ ($n = 1, 2, 3, \dots$). The F_n form a decreasing sequence of nonempty closed sets in X with $\text{diam } F_n \rightarrow 0$, but $\bigcap F_n = \emptyset$. There is no conflict with Theorem 6.3.1.

2. Condition (b) in the theorem on nested closed sets (6.3.1) requires that (i) the F_n are closed sets, (ii) $F_n \downarrow$, and (iii) $\text{diam } F_n \rightarrow 0$. If any of these three conditions is omitted, then the implication (a) \Rightarrow (b) fails.

{Hint: In $X = \mathbb{R}$ with the usual metric, contemplate the sequences (i) $F_n = (0, 1/n)$, (ii) $F_n = [n, n + 1/n]$, (iii) $F_n = [n, +\infty)$.}

3. In a compact metric space (which is complete, by 6.1.26) there is a shorter proof of the ‘nested closed sets property’.

{Hint: 6.1.5.}

4. (i) The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = x^2$ is continuous but not uniformly continuous (for the usual metric of \mathbb{R}).

(ii) For a bounded example, contemplate the function $f : (0, 1] \rightarrow \mathbb{R}$, $f(x) = \sin(1/x)$.

5. If $X = Y = \mathbb{R}$ with the usual metric and $f : \mathbb{R} \rightarrow \mathbb{R}$ is the function $f(x) = x^2$, then f satisfies condition (c) of 6.3.4 (because \mathbb{R} is complete and f is continuous) but f is not uniformly continuous.

6.4. L^1

Topology, measure theory and functional analysis grew up together in the first half of the 20th century, so it is not surprising that some of the most useful examples of metric spaces of analysis are based on integration over a measure space. This section is devoted to such an example, a complete metric space associated with the class of integrable functions relative to a measure space.

For the rest of the section, (X, \mathcal{S}, μ) is a fixed measure space; thus, X is a set, \mathcal{S} is a σ -algebra of subsets of X , and μ is a measure defined on the σ -algebra \mathcal{S} (2.4.12).

We begin by bringing complex-valued functions on board:

6.4.1. *Definition.* A function $f : X \rightarrow \mathbb{C}$ is said to be **measurable** (with respect to the σ -algebra \mathcal{S}) if its real and imaginary parts

$$\text{Re } f = \frac{1}{2}(f + \overline{f}), \quad \text{Im } f = \frac{1}{2i}(f - \overline{f})$$

(regarded as functions $X \rightarrow \mathbb{R}$) are measurable with respect to \mathcal{S} in the sense of Definition 4.1.3, in other words, if $f = g + ih$ with $g, h : X \rightarrow \mathbb{R}$ measurable in the sense of 4.1.3.

It is clear from the definition that the correspondence $g \mapsto g + 0i$ maps the set of all measurable functions $g : X \rightarrow \mathbb{R}$ onto the set of all measurable functions $f : X \rightarrow \mathbb{C}$ that are real-valued (that is, for which $f(X) \subset \mathbb{R}$) and that this correspondence preserves the algebraic operations (pointwise sums, products, and scalar multiples by real scalars).

6.4.2. Proposition. *If $f, g : X \rightarrow \mathbb{C}$ are measurable, $c \in \mathbb{C}$ and $\alpha > 0$ then the functions $f + g$, cf , fg $|f|^\alpha$ are also measurable.*

Proof. Write $f = f_1 + if_2$ and $g = g_1 + ig_2$ with f_1, f_2, g_1, g_2 real-valued, and suppose $c = a + ib$ with $a, b \in \mathbb{R}$. Then

$$\begin{aligned} f + g &= (f_1 + g_1) + i(f_2 + g_2), \\ cf &= (af_1 - bf_2) + i(af_2 + bf_1), \\ fg &= (f_1g_1 - f_2g_2) + i(f_1g_2 + f_2g_1), \\ |f|^\alpha &= (|f_1|^2 + |f_2|^2)^{\alpha/2}; \end{aligned}$$

the real and imaginary parts of the functions on the right side are measurable by Theorems 4.1.9 and 4.1.13. \diamond

In particular, the measurable complex-valued functions form a vector space over the field of complex numbers. As in the real-valued case, the pointwise limit of a sequence of measurable complex-valued functions is measurable:

6.4.3. Proposition. *If (f_n) is a sequence of measurable complex functions, $f : X \rightarrow \mathbb{C}$ and $f_n \rightarrow f$ pointwise, then the limit function f is also measurable.*

Proof. Since $\operatorname{Re} f_n \rightarrow \operatorname{Re} f$ and $\operatorname{Im} f_n \rightarrow \operatorname{Im} f$ pointwise, the assertion is immediate from Corollary 4.1.20. \diamond

Recall that $\mathcal{L}_{\mathbb{R}}^1 = \mathcal{L}_{\mathbb{R}}^1(X, \mathcal{S}, \mu)$ denotes the real vector space of real-valued functions $f : X \rightarrow \mathbb{R}$ that are integrable with respect to μ (4.4.7).

6.4.4. Definition. A complex function $f : X \rightarrow \mathbb{C}$ is said to be **integrable** with respect to the measure μ if its real and imaginary parts are integrable in the sense of 4.4.7, in other words, if $\operatorname{Re} f, \operatorname{Im} f \in \mathcal{L}_{\mathbb{R}}^1$. We write

$$\mathcal{L}_{\mathbb{C}}^1 = \mathcal{L}_{\mathbb{C}}^1(X, \mathcal{S}, \mu)$$

for the set of all such functions f , and the **integral** (with respect to μ) of such a function, denoted $\int f d\mu$, is defined by the formula

$$\int f d\mu = \int (\operatorname{Re} f) d\mu + i \int (\operatorname{Im} f) d\mu.$$

If $g \in \mathcal{L}_{\mathbb{R}}^1$ then the function $f : X \rightarrow \mathbb{C}$ defined by $f = g + 0i$ belongs to $\mathcal{L}_{\mathbb{C}}^1$, and the integral $\int f d\mu$ just defined coincides with $\int g d\mu$ as

defined in 4.4.9. Identifying g with $g + 0i$, we may regard $\mathcal{L}_{\mathbb{R}}^1$ as the set of all $f \in \mathcal{L}_{\mathbb{C}}^1$ with $f(X) \subset \mathbb{R}$.

6.4.5. Proposition. $\mathcal{L}_{\mathbb{C}}^1$ is a complex vector space and $f \mapsto \int f d\mu$ is a linear form on it.

Proof. This follows easily from the case of real-valued functions (4.4.11) and the formulas in the proof of Proposition 6.4.2. \diamond

6.4.6. Proposition. The following conditions on a complex function $f : X \rightarrow \mathbb{C}$ are equivalent:

(a) $f \in \mathcal{L}_{\mathbb{C}}^1$;

(b) f is measurable and $|f| \in \mathcal{L}_{\mathbb{R}}^1$.

Moreover, $|\int f d\mu| \leq \int |f| d\mu$ for every $f \in \mathcal{L}_{\mathbb{C}}^1$.

Proof. (a) \Rightarrow (b): Assuming $f \in \mathcal{L}_{\mathbb{C}}^1$, write $f = g + ih$ with $g, h \in \mathcal{L}_{\mathbb{R}}^1$ (6.4.4). Then f is measurable (6.4.1), therefore so is $|f|$ (6.4.2), and

$$|f| = (g^2 + h^2)^{1/2} \leq |g| + |h|$$

(to check the inequality, square both sides); since $|g| + |h|$ is integrable (4.4.11), so is $|f|$ (4.4.5).

To verify the asserted inequality, write $|\int f d\mu| = c \int f d\mu$ with $|c| = 1$ (if $\int f d\mu = 0$ let $c = 1$; if $\alpha = \int f d\mu \neq 0$ let $c = |\alpha|/\alpha$). Say $c = a + ib$ ($a, b \in \mathbb{R}$). Then

$$\begin{aligned} \left| \int f d\mu \right| &= \int c f d\mu = \int [(ag - bh) + i(ah + bg)] d\mu \\ &= \int (ag - bh) d\mu + i \int (ah + bg) d\mu; \end{aligned}$$

since the left side is real, necessarily $\int (ah + bg) d\mu = 0$ and

$$\begin{aligned} \left| \int f d\mu \right| &= \int (ag - bh) d\mu \\ &= \left| \int (ag - bh) d\mu \right| \\ &\leq \int |ag - bh| d\mu \end{aligned}$$

(cf. 4.4.12). Then

$$|ag - bh| \leq \sqrt{a^2 + b^2} \cdot \sqrt{g^2 + h^2} = |c| \cdot |f| = |f|$$

by Cauchy's inequality in \mathbb{R}^2 (or by 3.1.12 with $p = 2$), applied after evaluating the functions at each point of X ; combining the preceding inequalities, we have

$$\left| \int f d\mu \right| \leq \int |ag - bh| d\mu \leq \int |f| d\mu.$$

(b) \Rightarrow (a): With f as in (b), write $f = g + ih$, where $g = \operatorname{Re} f$ and $h = \operatorname{Im} f$. By hypothesis, g and h are measurable (6.4.1); moreover,

$$|g| \leq (g^2 + h^2)^{1/2} = |f| \in \mathcal{L}_{\mathbb{R}}^1,$$

therefore $g \in \mathcal{L}_{\mathbb{R}}^1$ (4.4.5, 4.4.12) and similarly $h \in \mathcal{L}_{\mathbb{R}}^1$, thus $f \in \mathcal{L}_{\mathbb{C}}^1$ (6.4.4). \diamond

In particular, $\mathcal{L}_{\mathbb{C}}^1$ is closed under the operation $f \mapsto |f|$.

6.4.7. *Definition.* For $f \in \mathcal{L}_{\mathbb{C}}^1$ we write

$$\|f\|_1 = \int |f| d\mu,$$

called the \mathcal{L}^1 -**norm** of f .

6.4.8. *Proposition.* If $f, g \in \mathcal{L}_{\mathbb{C}}^1$ and $c \in \mathbb{C}$, then

- (1) $\|f\|_1 \geq 0$,
- (2) $\|cf\|_1 = |c| \|f\|_1$,
- (3) $\|f + g\|_1 \leq \|f\|_1 + \|g\|_1$.

Proof. Properties (1) and (2) are clear from the definition, and (3) follows from integrating the inequality $|f + g| \leq |f| + |g|$. \diamond

6.4.9. *Definition.* A real-valued function $x \mapsto \|x\|$ defined on a (real or complex) vector space V is said to be a **seminorm** if it satisfies the conditions (1)–(3) of the above proposition (with c restricted to real scalars in the case of a real vector space). If, in addition, (4) $\|x\| > 0$ for every nonzero vector x , then the function $x \mapsto \|x\|$ is called a **norm** on V , $\|x\|$ is called the *norm* of the vector x , and the pair $(V, \|\cdot\|)$ — V equipped with the norm function $x \mapsto \|x\|$ —is called a **normed vector space** (or, briefly, a **normed space**).

6.4.10. *Remark.* If $x \mapsto \|x\|$ is a seminorm on the vector space V , then the formula $d(x, y) = \|x - y\|$ defines a *pseudometric* on V (3.1.4); in particular, the triangle inequality for d follows on applying property (3) of the preceding definition to the sum $x - y = (x - z) + (z - y)$. The correspondence $f \mapsto \|f\|_1$ of 6.4.7 is a seminorm on $\mathcal{L}_{\mathbb{C}}^1$ (6.4.8); it is not in general a norm, since $\|f\|_1 = 0$ whenever $f = 0$ a.e. (4.4.21), which does not in general imply that $f = 0$.

6.4.11. *Definition.* We write $\mathcal{N} = \mathcal{N}(X, \mathcal{S}, \mu)$ for the set of all *measurable* functions $f : X \rightarrow \mathbb{C}$ such that $f = 0$ a.e. (equivalently, $\operatorname{Re} f = 0$ a.e. and $\operatorname{Im} f = 0$ a.e.). Such functions are called **null functions** (relative to the given measure space); they are the integrable functions whose absolute value has integral 0:

6.4.12. Proposition. *The set \mathcal{N} of null functions is a linear subspace of $\mathcal{L}_{\mathbb{C}}^1$, and*

$$\mathcal{N} = \{f \in \mathcal{L}_{\mathbb{C}}^1 : \|f\|_1 = 0\}.$$

Proof. If $f \in \mathcal{N}$ then $|f|$ is measurable (6.4.2) and $|f| = 0$ a.e., therefore $|f|$ is integrable and $\int |f| d\mu = 0$ (4.4.21). Thus $f \in \mathcal{L}_{\mathbb{C}}^1$ (6.4.6) and $\|f\|_1 = 0$.

Conversely, if $f \in \mathcal{L}_{\mathbb{C}}^1$ and $\int |f| d\mu = 0$, then $|f| = 0$ a.e. (4.4.21), whence $f \in \mathcal{N}$. Thus the stated formula for \mathcal{N} is verified. In particular, \mathcal{N} contains the zero function.

If $f, g \in \mathcal{N}$ and $c \in \mathbb{C}$ then $\|f + g\|_1 \leq \|f\|_1 + \|g\|_1 = 0$ and $\|cf\|_1 = |c| \|f\|_1 = 0$, therefore $f + g, cf \in \mathcal{N}$; thus \mathcal{N} is a linear subspace of $\mathcal{L}_{\mathbb{C}}^1$. \diamond

In particular, for $f, g \in \mathcal{L}_{\mathbb{C}}^1$ we have

$$f = g \text{ a.e.} \Leftrightarrow f - g \in \mathcal{N} \Leftrightarrow \|f - g\|_1 = 0.$$

The vector space $\mathcal{L}_{\mathbb{C}}^1$, equipped with the pseudometric $(f, g) \mapsto \|f - g\|_1$ derived from the seminorm $f \mapsto \|f\|_1$ (cf. 6.4.10), is ‘complete’ in the following sense:

6.4.13. Theorem. *If (f_n) is a sequence in $\mathcal{L}_{\mathbb{C}}^1$ such that*

$$\|f_m - f_n\|_1 \rightarrow 0 \quad \text{as } m, n \rightarrow \infty$$

(such a sequence is said to be ‘Cauchy in mean’ or ‘fundamental in mean’), then there exists a function $f \in \mathcal{L}_{\mathbb{C}}^1$ such that

$$\|f_n - f\|_1 \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

(and one says that $f_n \rightarrow f$ ‘in mean’). If also $g \in \mathcal{L}_{\mathbb{C}}^1$ and $\|f_n - g\|_1 \rightarrow 0$, then $f = g$ a.e. (so to speak, f is ‘a.e. unique’).

Proof. Passing to a subsequence (cf. the proof of 6.1.26), we can suppose that

$$\|f_{n+1} - f_n\|_1 \leq 2^{-n} \quad (n = 1, 2, 3, \dots).$$

Write $f_0 = 0$ and, for every positive integer n , define

$$h_n = \sum_{k=1}^n |f_k - f_{k-1}| = |f_1| + \sum_{k=2}^n |f_k - f_{k-1}|.$$

The functions h_n are integrable (§4.4) and form an increasing sequence,

$h_n \uparrow$; moreover, the sequence of integrals $\int h_n d\mu$ is bounded,

$$\begin{aligned} \int h_n d\mu &= \|f_1\|_1 + \sum_{k=2}^n \|f_k - f_{k-1}\|_1 \\ &= \|f_1\|_1 + \sum_{k=1}^{n-1} \|f_{k+1} - f_k\|_1 \\ &\leq \|f_1\|_1 + \sum_{k=1}^{n-1} 2^{-k} < \|f_1\|_1 + 1, \end{aligned}$$

so by the monotone convergence theorem (4.5.3) there exists an integrable function h such that $h_n \uparrow h$ a.e. Let E be a null set on whose complement $h_n(x) \uparrow h(x)$; replacing f_n by $\varphi_{E^c} f_n$, we can suppose that $h_n(x) \uparrow h(x)$ for every point x of X . Note that

$$|f_n| = \left| \sum_{k=1}^n (f_k - f_{k-1}) \right| \leq \sum_{k=1}^n |f_k - f_{k-1}| = h_n \leq h$$

for all n ; we will show that the sequence (f_n) is pointwise convergent.

For every positive integer n , write $s_n = f_n - f_{n-1}$; then, for every point $x \in X$,

$$\sum_{k=1}^n |s_k(x)| = h_n(x) \leq h(x),$$

so the series $\sum_{k=1}^{\infty} s_k(x)$ is convergent (absolutely). Define a function $f: X \rightarrow \mathbb{C}$ by the formula

$$\begin{aligned} f(x) &= \sum_{k=1}^{\infty} s_k(x) = \lim_{n \rightarrow \infty} \sum_{k=1}^n s_k(x) \\ &= \lim_{n \rightarrow \infty} \sum_{k=1}^n [f_k(x) - f_{k-1}(x)] = \lim_{n \rightarrow \infty} f_n(x), \end{aligned}$$

thus $f_n \rightarrow f$ pointwise on X ; in particular, f is measurable (6.4.3) and

$$|f(x)| = \left| \sum_{k=1}^{\infty} s_k(x) \right| \leq \sum_{k=1}^{\infty} |s_k(x)| \leq h(x)$$

for all $x \in X$. It follows from $|f| \leq h \in \mathcal{L}_{\mathbb{R}}^1$ that $|f| \in \mathcal{L}_{\mathbb{R}}^1$ (4.4.19), therefore $f \in \mathcal{L}_{\mathbb{C}}^1$ (6.4.6); and, for every positive integer n ,

$$|f_n - f| \leq |f_n| + |f| \leq |f_n| + h \leq 2h \in \mathcal{L}_{\mathbb{R}}^1;$$

since $|f_n - f| \rightarrow 0$ pointwise on X , it follows from the dominated convergence theorem (4.5.4) that $\int |f_n - f| d\mu \rightarrow 0$. This proves the existence of a function $f \in \mathcal{L}_{\mathbb{C}}^1$ such that $\|f_n - f\|_1 \rightarrow 0$.

If also $g \in \mathcal{L}_\mathbb{C}^1$ and $\|f_n - g\|_1 \rightarrow 0$, then

$$\|f - g\|_1 \leq \|f - f_n\|_1 + \|f_n - g\|_1 \rightarrow 0,$$

therefore $\|f - g\|_1 = 0$, whence $f - g = 0$ a.e. (6.4.12), that is, $f = g$ a.e. \diamond

The message of the preceding theorem is that the vector space $\mathcal{L}_\mathbb{C}^1$, equipped with the pseudometric $d(f, g) = \|f - g\|_1$, is 'complete' in the sense that if a sequence (f_n) in $\mathcal{L}_\mathbb{C}^1$ satisfies the Cauchy condition $d(f_m, f_n) \rightarrow 0$, then $d(f_n, f) \rightarrow 0$ for a suitable $f \in \mathcal{L}_\mathbb{C}^1$; the 'limit function' f is 'a.e. unique', in the sense that if also $g \in \mathcal{L}_\mathbb{C}^1$ and $d(f_n, g) \rightarrow 0$, then the difference function $f - g$ belongs to the linear subspace \mathcal{N} of null functions in $\mathcal{L}_\mathbb{C}^1$ (6.4.12). To pass from the pseudometric space $(\mathcal{L}_\mathbb{C}^1, d)$ to a complete metric space, we need only pass to the quotient vector space $\mathcal{L}_\mathbb{C}^1/\mathcal{N}$:

6.4.14. *Definition.* The quotient vector space $\mathcal{L}_\mathbb{C}^1/\mathcal{N}$ is denoted

$$L_\mathbb{C}^1(X, \mathcal{S}, \mu),$$

briefly $L_\mathbb{C}^1$. If $u \in L_\mathbb{C}^1$, say $u = f + \mathcal{N}$, we write also $u = \dot{f}$; if also $u = \dot{g}$ —that is, if $f - g \in \mathcal{N}$ —then $f = g$ a.e. (6.4.11), so we may define

$$\|u\|_1 = \|f\|_1,$$

which depends only on the coset u and not on the particular function f that represents it. Thus, by definition, $\|\dot{f}\|_1 = \|f\|_1$.

6.4.15. *Proposition.* The correspondence $u \mapsto \|u\|_1$ of 6.4.14 defines a norm on $L_\mathbb{C}^1$; that is, if $u, v \in L_\mathbb{C}^1$ and $c \in \mathbb{C}$, then

- (1) $\|u\|_1 \geq 0$,
- (2) $\|cu\|_1 = |c| \|u\|_1$,
- (3) $\|u + v\|_1 \leq \|u\|_1 + \|v\|_1$, and
- (4) $\|u\|_1 > 0 \Leftrightarrow u \neq 0$.

Proof. (1)–(3) are immediate from 6.4.8.

(4) Say $u = \dot{f}$; then $\|u\|_1 = \|f\|_1$, so

$$\|u\|_1 = 0 \Leftrightarrow \|f\|_1 = 0 \Leftrightarrow f \in \mathcal{N} \text{ (6.4.12)} \Leftrightarrow \dot{f} = 0;$$

thus $\|u\|_1 = 0 \Leftrightarrow u = 0$, whence (4). \diamond

In other words, $(L_\mathbb{C}^1, \|\cdot\|_1)$ is a normed space (6.4.9). By the same arguments as in §3.1, the formula $d(u, v) = \|u - v\|_1$ defines a metric on $L_\mathbb{C}^1$; in fact, it is a complete metric:

6.4.16. *Proposition.* If (u_n) is a sequence in $L_\mathbb{C}^1$ with

$$\|u_m - u_n\|_1 \rightarrow 0 \text{ as } m, n \rightarrow \infty,$$

then there exists $u \in L_\mathbb{C}^1$ such that $\|u_n - u\|_1 \rightarrow 0$.

Proof. Immediate from 6.4.13. \diamond

6.4.17. *Definition.* A (real or complex) normed space that is complete for the metric $(x, y) \mapsto \|x - y\|$ deduced from the norm is called a **Banach space**.

Propositions 6.4.15 and 6.4.16 can be expressed as follows:

6.4.18. *Theorem.* For every measure space (X, \mathcal{S}, μ) , the quotient vector space

$$L_{\mathbb{C}}^1 = \mathcal{L}_{\mathbb{C}}^1(X, \mathcal{S}, \mu) / \mathcal{N}(X, \mathcal{S}, \mu)$$

is a Banach space for the norm $u \mapsto \|u\|_1$ of 6.4.14.

The superscript in the symbol $L_{\mathbb{C}}^1$ alludes to the fact that it is the first power of $|f|$ that is being integrated when one defines $\|\dot{f}\|_1 = \|f\|_1 = \int |f| d\mu$. In §6.7 below, consideration of $|f|^p$, $p > 1$, leads to Banach spaces $L_{\mathbb{C}}^p$ analogous to (indeed, generalizing) the finite-dimensional “Minkowski spaces” of 3.1.11.

We conclude this section with the complex analogue of a result in §4.7:

6.4.19. *Proposition.* Let $\mathcal{A} \subset \mathcal{S}$ be an algebra of sets such that the σ -algebra generated by \mathcal{A} is \mathcal{S} . If $f, g \in \mathcal{L}_{\mathbb{C}}^1$ and if

$$\int_A f d\mu = \int_A g d\mu \quad \text{for all } A \in \mathcal{A},$$

then $f = g$ a.e.

Proof. If $E \in \mathcal{S}$ then $\varphi_E f \in \mathcal{L}_{\mathbb{C}}^1$ follows from the case of real-valued functions (4.4.18) and Definition 6.4.4; following 4.4.22, one defines $\int_E f d\mu$ to be $\int \varphi_E f d\mu$.

Assuming $f, g \in \mathcal{L}_{\mathbb{C}}^1$ satisfy the condition of the hypothesis, let $h = f - g$; our assumption is that

$$(*) \quad \int_A h d\mu = 0 \quad \text{for all } A \in \mathcal{A},$$

and we seek to show that $h = 0$ a.e. Writing $h = u + iv$ with $u, v \in \mathcal{L}_{\mathbb{R}}^1$, it is clear from Definition 6.4.4 that it suffices to consider the case that h is real-valued; but then $h = 0$ a.e. by Corollary 4.7.4. \diamond

Exercises

1. If V is a vector space (real or complex) and if $x \mapsto \|x\|$ is a seminorm on V (6.4.9), then the set $N = \{x \in V : \|x\| = 0\}$ is a linear subspace

of V and the formula $\|x + N\| = \|x\|$ defines a *norm* on the quotient vector space V/N .

2. If f and g are measurable complex-valued functions on a measurable space (X, \mathcal{S}) , then the function h defined by

$$h(x) = \begin{cases} \frac{f(x)}{g(x)} & \text{when } g(x) \neq 0 \\ 0 & \text{when } g(x) = 0 \end{cases}$$

is also measurable. More generally, if r is any measurable complex-valued function on X , one can require that $h(x) = r(x)$ whenever $g(x) = 0$.

{Hint: § 4.1, Exercise 3}

3. A measurable complex-valued function f can be written as $f = u|f|$, where u is measurable and $|u(x)| = 1$ for all x .

{Hint: Apply Exercise 2 with $g = |f|$ and $r = 1$.}

4. The set $\mathcal{BV}[a, b]$ of all functions $f : [a, b] \rightarrow \mathbb{R}$ of bounded variation, equipped with the pointwise linear operations, is a (real) Banach space for the norm $\|f\| = |f(a)| + V_a^b f$, as well as for the norm $\|f\|' = \|f\|_\infty + V_a^b f$ (§5.1, Exercise 4).

6.5. Real and Complex Measures

A *measure* is a function $\mu : \mathcal{S} \rightarrow [0, +\infty]$, defined on a σ -algebra \mathcal{S} , that is countably additive and vanishes at the empty set (2.4.12). In this section we consider the analogues for set functions with values in \mathbb{R} or in \mathbb{C} ; the slightly more delicate case of values in the extended reals $\overline{\mathbb{R}}$ is deferred until the final chapter, the delicacy being that the values $+\infty$ and $-\infty$ cannot both be taken on by a particular 'extended-real-valued measure' (§9.1).

6.5.1. *Definition.* Let \mathcal{S} be a σ -algebra of subsets of a set X . A **complex measure** on \mathcal{S} is a function $\nu : \mathcal{S} \rightarrow \mathbb{C}$ that is *countably additive* in the sense that

$$\nu \left(\bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} \nu(E_n)$$

whenever (E_n) is a sequence of pairwise disjoint sets in \mathcal{S} . If, moreover, ν is real-valued, it is called a **real measure** (or, as in §4.8, a **finite signed measure**) on \mathcal{S} .

6.5.2. *Remarks.* In the following remarks, ν, μ, \dots are complex measures on a σ -algebra \mathcal{S} .

1. A complex measure on \mathcal{S} is an element of the vector space $\mathcal{F}(\mathcal{S}, \mathbb{C})$ of complex-valued functions on \mathcal{S} , whence the possibility of performing

linear operations on complex measures. Since the sum and scalar multiples of complex measures are themselves complex measures (by the properties of term-by-term sums and scalar multiples of convergent series of complex numbers), the complex measures on \mathcal{S} form a linear subspace of $\mathcal{F}(\mathcal{S}, \mathbb{C})$; for example, if μ and ν are complex measures on \mathcal{S} , (E_n) is a sequence of pairwise disjoint sets in \mathcal{S} , and $E = \bigcup_{n=1}^{\infty} E_n$, then

$$\begin{aligned} (\mu + \nu)(E) &= \mu(E) + \nu(E) \\ &= \sum_{n=1}^{\infty} \mu(E_n) + \sum_{n=1}^{\infty} \nu(E_n) \\ &= \sum_{n=1}^{\infty} [\mu(E_n) + \nu(E_n)] \\ &= \sum_{n=1}^{\infty} (\mu + \nu)(E_n) \end{aligned}$$

(in particular, the last series is convergent¹), therefore $\mu + \nu$ is a complex measure.

2. If ν is a complex measure on \mathcal{S} , then so is the complex conjugate function $\bar{\nu}$, defined by

$$\bar{\nu}(E) = \overline{\nu(E)} \quad (E \in \mathcal{S});$$

the countable additivity of $\bar{\nu}$ follows from that of ν and from the continuity of complex conjugation in \mathbb{C} . It follows that ν is uniquely expressible as a linear combination $\nu = \rho + i\sigma$ of real measures ρ and σ , namely

$$\rho = \frac{1}{2}(\nu + \bar{\nu}) \quad \text{and} \quad \sigma = \frac{1}{2i}(\nu - \bar{\nu}),$$

called the real and imaginary parts of ν .

3. If μ and ν are (positive) finite measures on \mathcal{S} , then $\rho = \mu - \nu$ is a real measure on \mathcal{S} . Conversely, every real measure is a difference of positive measures (4.8.8), but not necessarily uniquely since, for example, $\mu - \nu = 2\mu - (\mu + \nu)$.

4. When ν is a real measure, the convergent series on the right in Definition 6.5.1 is *absolutely* convergent (it is 'commutatively convergent', since $\bigcup_{n=1}^{\infty} E_n$ is invariant under every permutation of the indices²). In view of Remark 2, the same is true for every complex measure ν .

5. $\nu(\emptyset) = 0$ (let $E_n = \emptyset$ for all n).

6. ν is finitely additive (by Remark 5 and countable additivity); it follows that ν is *subtractive*, that is, $\nu(F - E) = \nu(F) - \nu(E)$ when $F \supset E$, as one sees by applying ν to the disjoint union $(F - E) \cup E = F$ (cf. 2.6.1).

¹ Cf. *First course*, p. 183, Theorem 10.2.1.

² Cf. E. Landau, *Differential and integral calculus* [Chelsea, New York, 1951], p. 158, Theorem 217; W. Rudin, *Principles of mathematical analysis* [3rd edn., McGraw-Hill, New York, 1976], p. 76, Theorem 3.54.

7. If $E_n \uparrow E$ then $\nu(E_n) \rightarrow \nu(E)$, as one sees by applying ν to the countable disjoint union

$$E = E_1 \cup (E_2 - E_1) \cup (E_3 - E_2) \cup \dots;$$

similarly, $E_n \downarrow E$ implies $\nu(E_n) \rightarrow \nu(E)$ (cf. 2.6.2, 2.6.3).

Two complex measures on \mathcal{S} that agree on a generating subalgebra are identical:

6.5.3. Theorem. Let \mathcal{A} be an algebra of subsets of a set X (2.4.1), and let \mathcal{S} be the σ -algebra generated by \mathcal{A} (2.4.4). If ν_1 and ν_2 are complex measures on \mathcal{S} such that

$$\nu_1(E) = \nu_2(E) \quad \text{for all } E \in \mathcal{A},$$

then $\nu_1 = \nu_2$ on \mathcal{S} .

Proof. (Cf. 4.6.7.) Let $\mathcal{T} = \{E \in \mathcal{S} : \nu_1(E) = \nu_2(E)\}$. By assumption $\mathcal{A} \subset \mathcal{T}$, and \mathcal{T} is a monotone class by the preceding Remark 7, therefore $\mathcal{S} \subset \mathcal{T}$ by the Lemma on monotone classes (4.6.6). \diamond

6.5.4. Corollary. Let $[a, b]$ be a closed interval in \mathbb{R} , $a < b$, and let \mathcal{B}_0 be the set of all Borel sets in \mathbb{R} that are contained in $[a, b]$. If ν_1 and ν_2 are complex measures on \mathcal{B}_0 such that

$$\nu_1([a, x]) = \nu_2([a, x]) \quad \text{for all } x \in [a, b],$$

then $\nu_1 = \nu_2$ on \mathcal{B}_0 .

Proof. Let \mathcal{C}_0 be the set of all subintervals of $[a, b]$, and let \mathcal{A}_0 be the algebra of subsets of $[a, b]$ generated by \mathcal{C}_0 . We know that $\mathcal{C}_0 \subset \mathcal{A}_0 \subset \mathcal{B}_0$ and that \mathcal{B}_0 is the σ -algebra of subsets of $[a, b]$ generated by \mathcal{C}_0 (noted in Example 4.7.5), therefore \mathcal{B}_0 is also the σ -algebra generated by \mathcal{A}_0 ; in view of the preceding theorem, we need only show that $\nu_1 = \nu_2$ on \mathcal{A}_0 , and, since \mathcal{A}_0 is the set of all finite disjoint unions of elements of \mathcal{C}_0 (4.6.2), it will suffice (by the additivity of the ν_i) to show that $\nu_1(I) = \nu_2(I)$ for every subinterval I of $[a, b]$.

case 1: $I = (c, d]$, where $a \leq c \leq d \leq b$.

Then $I = [a, d] - [a, c]$, and $\nu_1(I) = \nu_2(I)$ follows from the subtractivity of the ν_i (Remark 6 of 6.5.2).

case 2: $I = [a, d)$, where $a < d \leq b$.

Let $a < c_n < d$ with $c_n \uparrow d$. Then $[a, c_n] \uparrow [a, d)$, and $\nu_1(I) = \nu_2(I)$ follows from Remark 7 of 6.5.2.

case 3: I a singleton.

If $I = \{a\}$ then $I = [a, a]$ and $\nu_1(I) = \nu_2(I)$ by hypothesis, whereas if $I = \{d\}$ with $a < d \leq b$ then $\nu_1(I) = \nu_2(I)$ follows from the formula

$$\{d\} = [a, d] - [a, d)$$

and case 2.

The remaining cases then follow from the formulas

$$\begin{aligned}(c, d) &= (c, d] - \{d\}, \\ [c, d] &= \{c\} \cup (c, d), \\ [c, d] &= [c, d) \cup \{d\}. \quad \diamond\end{aligned}$$

6.5.5. Corollary. Let \mathcal{B}_0 be as in the preceding corollary, let μ be any measure on \mathcal{B}_0 , and suppose $f \in \mathcal{L}_{\mathbb{C}}^1(\mu)$ is such that

$$\int_a^x f d\mu = 0 \quad \text{for all } x \in [a, b],$$

where, by definition, $\int_a^x f d\mu = \int \varphi_{[a, x]} f d\mu$. Then $f = 0$ a.e. (relative to μ).

Proof. Since $f = g + ih$ with $g, h \in \mathcal{L}_{\mathbb{R}}^1(\mu)$, we can clearly suppose that $f \in \mathcal{L}_{\mathbb{R}}^1(\mu)$; writing $\nu = f \cdot \mu$ (4.7.1), we know ν is a real measure on \mathcal{S} (4.7.3) and, by assumption, $\nu([a, x]) = 0$ for all $x \in [a, b]$. By the preceding corollary, $\nu = 0$ on \mathcal{B}_0 , therefore $f = 0$ a.e. (4.7.4). \diamond

Exercises

1. (i) If ρ is a real measure on the σ -algebra \mathcal{S} , then there exist a finite measure μ on \mathcal{S} and a function $f \in \mathcal{L}_{\mathbb{R}}^1(\mu)$ such that $\rho = f \cdot \mu$ (cf. 4.7.1).

(ii) Extend (i) to complex measures.

{Hint: (i) Write ρ as a difference $\rho = \mu_1 - \mu_2$ of finite measures, let $\mu = \mu_1 + \mu_2$, and apply the Radon-Nikodym theorem (4.8.11) to μ_1, μ and to μ_2, μ .}

2. Let (X, \mathcal{S}, μ) be a measure space. A complex measure ν on \mathcal{S} is said to be *absolutely continuous* with respect to μ , written $\nu \ll \mu$, in case ν vanishes on the null sets for μ , that is,

$$E \in \mathcal{S}, \mu(E) = 0 \quad \Rightarrow \quad \nu(E) = 0$$

(cf. 4.8.6).

(i) If $\nu = \nu_1 + i\nu_2$ with ν_1 and ν_2 real measures on \mathcal{S} , then

$$\nu \ll \mu \quad \Leftrightarrow \quad \nu_1 \ll \mu \quad \text{and} \quad \nu_2 \ll \mu.$$

(ii) If $f \in \mathcal{L}_{\mathbb{C}}^1(\mu)$ then the set function $f \cdot \mu : \mathcal{S} \rightarrow \mathbb{C}$ defined by $(f \cdot \mu)(E) = \int \varphi_E f d\mu$ is a complex measure on \mathcal{S} such that $f \cdot \mu \ll \mu$.

(iii) If the measure μ is finite and if ν is a complex measure on \mathcal{S} such that $\nu \ll \mu$, then there exists a function $f \in \mathcal{L}_{\mathbb{C}}^1(\mu)$ such that $\nu = f \cdot \mu$.

{Hint: Cf. 4.8.12.}

6.6. L^∞

Throughout this section, (X, \mathcal{S}, μ) is a fixed measure space (2.4.12), eventually specialized to the example $([a, b], \mathcal{M}_0, \lambda_0)$ of Lebesgue measure on a closed interval (4.7.6).

6.6.1. *Definition.* A function $f : X \rightarrow \mathbb{C}$ is said to be **essentially bounded** (with respect to μ) if there exists a real number $M \geq 0$ such that $|f| \leq M$ a.e. (with respect to μ). Such a number M is called an **essential bound** for f (more aptly, for $|f|$).

6.6.2. *Lemma.* Every essentially bounded function has a smallest essential bound.

Proof. Suppose $f : X \rightarrow \mathbb{C}$ is essentially bounded and let S be the set of all essential bounds for f ,

$$S = \{M \geq 0 : |f| \leq M \text{ a.e.}\};$$

by assumption, $S \neq \emptyset$. Let $M = \inf S$; it will suffice to show that $M \in S$.

Choose a sequence $M_n \in S$ with $M_n \rightarrow M$. For each index n , let $E_n \in \mathcal{S}$ be a set of measure zero such that $|f| \leq M_n$ on $E'_n = X - E_n$. Then $E = \bigcup_{n=1}^{\infty} E_n$ is a measurable set of measure zero, and

$$x \in E' = \bigcap_{n=1}^{\infty} E'_n \Rightarrow |f(x)| \leq M_n \text{ for all } n \Rightarrow |f(x)| \leq M,$$

thus M is an essential bound for f , in other words, $M \in S$. \diamond

6.6.3. *Definition.* If $f : X \rightarrow \mathbb{C}$ is essentially bounded, the smallest essential bound for f (6.6.2) is denoted $\|f\|_\infty$ and is called the **essential supremum** of f (more aptly, of $|f|$).

CAUTION: In another context, $\|f\|_\infty$ stands for the supremum of $|f|$ (cf. 3.1.10).

6.6.4. *Definition.* The set of all functions $f : X \rightarrow \mathbb{C}$ that are measurable (with respect to \mathcal{S}) and essentially bounded (with respect to μ) is denoted

$$\mathcal{L}_\mathbb{C}^\infty(X, \mathcal{S}, \mu),$$

briefly $\mathcal{L}_\mathbb{C}^\infty$ or, when it is necessary to indicate the measure in question, $\mathcal{L}_\mathbb{C}^\infty(\mu)$.

As in 6.4.11, we write $\mathcal{N} = \mathcal{N}(X, \mathcal{S}, \mu)$ for the set of all measurable functions $f : X \rightarrow \mathbb{C}$ such that $f = 0$ a.e.

6.6.5. *Proposition.* $\mathcal{L}_\mathbb{C}^\infty$ is a subalgebra of the algebra $\mathcal{F}(X, \mathbb{C})$ of all complex-valued functions on X , and \mathcal{N} is an ideal of $\mathcal{L}_\mathbb{C}^\infty$. For all $f, g \in \mathcal{L}_\mathbb{C}^\infty$ and $c \in \mathbb{C}$,

$$(1) \|f\|_\infty \geq 0,$$

$$(2) \|cf\|_\infty = |c| \|f\|_\infty,$$

$$(3) \|f + g\|_\infty \leq \|f\|_\infty + \|g\|_\infty,$$

$$(4) \|f\|_\infty = 0 \Leftrightarrow f \in \mathcal{N},$$

$$(5) \|fg\|_\infty \leq \|f\|_\infty \|g\|_\infty.$$

In particular, the mapping $f \mapsto \|f\|_\infty$ is a seminorm on the complex vector space $\mathcal{L}_\mathbb{C}^\infty$.

Proof. The algebra operations in $\mathcal{F}(X, \mathbb{C})$ are the pointwise operations; for example, $(fg)(x) = f(x)g(x)$ for all $x \in X$. If $f, g \in \mathcal{L}_\mathbb{C}^\infty$ and $c \in \mathbb{C}$, then the functions $f + g$, cf and fg are measurable by Proposition 6.4.2.

(1) Obvious from Definition 6.6.3.

(2) If $c = 0$ then $cf = 0 \in \mathcal{L}_\mathbb{C}^\infty$ and the equality is obvious. Suppose $c \neq 0$. Then

$$|cf(x)| = |c| |f(x)| \leq |c| \|f\|_\infty \text{ a.e.}$$

(because $\|f\|_\infty$ is an essential bound for f), therefore $cf \in \mathcal{L}_\mathbb{C}^\infty$ and $\|cf\|_\infty \leq |c| \|f\|_\infty$. It follows that

$$\|f\|_\infty = \|c^{-1}(cf)\|_\infty \leq |c^{-1}| \|cf\|_\infty,$$

therefore $|c| \|f\|_\infty \leq \|cf\|_\infty$.

(3) $|f(x) + g(x)| \leq |f(x)| + |g(x)| \leq \|f\|_\infty + \|g\|_\infty$ a.e., therefore $f + g \in \mathcal{L}_\mathbb{C}^\infty$ and $\|f + g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$.

The message of (1)–(3) is that $f \mapsto \|f\|_\infty$ is a seminorm on $\mathcal{L}_\mathbb{C}^\infty$ (6.4.9).

(4) If $f \in \mathcal{N}$ then $|f| = 0$ a.e., therefore $|f| \leq 0$ a.e., whence $\|f\|_\infty \leq 0$; in view of (1), $\|f\|_\infty = 0$. This shows that $\mathcal{N} \subset \mathcal{L}_\mathbb{C}^\infty$ (as a linear subspace).

Conversely, if $f \in \mathcal{L}_\mathbb{C}^\infty$ and $\|f\|_\infty = 0$ then $|f| \leq 0$ a.e. (6.6.3), therefore $f = 0$ a.e., thus $f \in \mathcal{N}$.

(5) $|f(x)g(x)| = |f(x)| |g(x)| \leq \|f\|_\infty \|g\|_\infty$ a.e., therefore $fg \in \mathcal{L}_\mathbb{C}^\infty$ and $\|fg\|_\infty \leq \|f\|_\infty \|g\|_\infty$.

From (4) and (5) we see that if $f \in \mathcal{L}_\mathbb{C}^\infty$ and $g \in \mathcal{N}$ then $fg \in \mathcal{N}$, therefore \mathcal{N} is an ideal of $\mathcal{L}_\mathbb{C}^\infty$. \diamond

6.6.6. *Definition.* The quotient algebra $\mathcal{L}_\mathbb{C}^\infty / \mathcal{N}$ is denoted

$$L_\mathbb{C}^\infty(X, \mathcal{S}, \mu),$$

briefly $L_\mathbb{C}^\infty$ or $L_\mathbb{C}^\infty(\mu)$, the operations on cosets being given by the formulas

$$\begin{aligned} (f + \mathcal{N}) + (g + \mathcal{N}) &= (f + g) + \mathcal{N} \\ c(f + \mathcal{N}) &= cf + \mathcal{N} \\ (f + \mathcal{N})(g + \mathcal{N}) &= fg + \mathcal{N} \end{aligned},$$

for all $f, g \in \mathcal{L}_\mathbb{C}^\infty$ and $c \in \mathbb{C}$. If $f \in \mathcal{L}_\mathbb{C}^\infty$ we write

$$\dot{f} = f + \mathcal{N}$$

for the image of f under the quotient homomorphism $\mathcal{L}_\mathbb{C}^\infty \rightarrow L_\mathbb{C}^\infty$.

If $u \in L_{\mathbb{C}}^\infty$, say $u = \dot{f} = \dot{g}$, then $f - g \in \mathcal{N}$ and therefore $\|f - g\|_\infty = 0$; since

$$\|f\|_\infty = \|(f - g) + g\|_\infty \leq \|f - g\|_\infty + \|g\|_\infty = \|g\|_\infty$$

and similarly $\|g\|_\infty \leq \|f\|_\infty$, so that $\|f\|_\infty = \|g\|_\infty$, it follows that the definition

$$\|u\|_\infty = \|f\|_\infty$$

is independent of the particular representative f of the coset u .

It then follows from Proposition 6.6.5 (by the same argument as in 6.4.15) that $u \mapsto \|u\|_\infty$ is a *norm* on the complex vector space $L_{\mathbb{C}}^\infty$. In fact:

6.6.7. Theorem. $L_{\mathbb{C}}^\infty(\mu)$ is a Banach space for the norm $u \mapsto \|u\|_\infty$.

Proof. The problem is to show that every Cauchy sequence in $L_{\mathbb{C}}^\infty$ is convergent (6.4.17). The crux of the matter is as follows: assuming (f_n) is a sequence of functions in $\mathcal{L}_{\mathbb{C}}^\infty$ such that $\|f_m - f_n\|_\infty \rightarrow 0$ as $m, n \rightarrow \infty$, we seek a function $f \in \mathcal{L}_{\mathbb{C}}^\infty$ such that $\|f_n - f\|_\infty \rightarrow 0$.

For each pair of indices $m, n \in \mathbb{P}$ let $E_{mn} \in \mathcal{S}$ be a set of measure zero such that

$$|f_m(x) - f_n(x)| \leq \|f_m - f_n\|_\infty \quad \text{for all } x \in E'_{mn};$$

then $E = \bigcup_{m,n=1}^\infty E_{mn}$ has measure zero and, for each $x \in E'$, the inequalities

$$|f_m(x) - f_n(x)| \leq \|f_m - f_n\|_\infty$$

show that $(f_n(x))$ is a Cauchy sequence of complex numbers. Define $f: X \rightarrow \mathbb{C}$ by

$$f(x) = \begin{cases} \lim f_n(x) & \text{for } x \in E' \\ 0 & \text{for } x \in E. \end{cases}$$

Since $\varphi_{E'} f_n \rightarrow f$ pointwise on X , the limit function f is measurable (4.1.20).

For all $x \in E'$ and for every pair of indices m, n ,

$$\begin{aligned} (*) \quad |f(x) - f_n(x)| &\leq |f(x) - f_m(x)| + |f_m(x) - f_n(x)| \\ &\leq |f(x) - f_m(x)| + \|f_m - f_n\|_\infty. \end{aligned}$$

Given any $\epsilon > 0$, choose an index N such that

$$m, n \geq N \Rightarrow \|f_m - f_n\|_\infty \leq \epsilon.$$

Fix a pair of indices $m, n \geq N$. For each $x \in E'$, it follows from (*) that

$$|f(x) - f_n(x)| \leq |f(x) - f_m(x)| + \epsilon;$$

keeping n fixed and letting $m \rightarrow \infty$, we have

$$|f(x) - f_n(x)| \leq 0 + \epsilon.$$

This shows that $f - f_n \in \mathcal{L}_\mathbb{C}^\infty$, hence also

$$f = (f - f_n) + f_n \in \mathcal{L}_\mathbb{C}^\infty,$$

and that $\|f - f_n\|_\infty \leq \epsilon$; since the inequality holds for all $n \geq N$, we have shown that $\|f - f_n\|_\infty \rightarrow 0$. \diamond

The next results explore the relation between the Banach spaces $L_\mathbb{C}^\infty(\mu)$ and $L_\mathbb{C}^1(\mu)$.

6.6.8. Lemma. *If f is integrable and g is essentially bounded and measurable, then fg is integrable and $\|fg\|_1 \leq \|f\|_1 \|g\|_\infty$.*

Proof. We are assuming $f \in \mathcal{L}_\mathbb{C}^1$ and $g \in \mathcal{L}_\mathbb{C}^\infty$. Since fg is measurable (6.4.2), $|f|$ is integrable (6.4.6) and

$$|fg| \leq |f| \|g\|_\infty \text{ a.e.,}$$

it follows that $|fg|$ is integrable (4.4.20), therefore so is fg (6.4.6), and

$$\|fg\|_1 = \int |fg| d\mu \leq \|g\|_\infty \int |f| d\mu = \|g\|_\infty \|f\|_1. \diamond$$

6.6.9. Theorem. *If $g \in \mathcal{L}_\mathbb{C}^\infty$ then the formula*

$$T\dot{f} = \int fg d\mu \quad (f \in \mathcal{L}_\mathbb{C}^1)$$

defines a continuous linear form on the Banach space $L_\mathbb{C}^1$.

Proof. It is clear from the lemma that T is a well-defined linear form on $L_\mathbb{C}^1$, where “well-defined” means that for $u = \dot{f} \in L_\mathbb{C}^1$, Tu depends only on the equivalence class u and not on the particular function f representing u . Moreover, by Proposition 6.4.6 and the lemma,

$$|T\dot{f}| = \left| \int fg d\mu \right| \leq \int |fg| d\mu \leq \|f\|_1 \|g\|_\infty,$$

thus $|Tu| \leq \|u\|_1 \|g\|_\infty$ for all $u \in L_\mathbb{C}^1$. Thus, for all $u, v \in L_\mathbb{C}^1$,

$$|Tv - Tu| = |T(v - u)| \leq \|v - u\|_1 \|g\|_\infty,$$

whence the continuity of T :

$$\|u_n - u\|_1 \rightarrow 0 \Rightarrow \|Tu_n - Tu\| \rightarrow 0. \diamond$$

Is the converse of the theorem true, that is, is *every* continuous linear form T on $L_\mathbb{C}^1$ given by the formula of the theorem for a suitable function $g \in \mathcal{L}_\mathbb{C}^\infty$? The answer depends on the measure space, and may be “no”.¹ For Lebesgue measure, the answer is “yes”; for simplicity, we limit the following

¹ Cf. E. Hewitt and K. Stromberg, *Real and abstract analysis* [Springer, New York, 1965], p.349, (20.17).

discussion to a closed interval $[a, b]$, $a < b$, and to the real Banach space $L^1_{\mathbb{R}}[a, b]$, with notations as in 4.7.6:

$$\begin{aligned}\lambda &= \text{Lebesgue measure on } \mathbb{R} \\ \mathcal{M} &= \{E \subset \mathbb{R} : E \text{ Lebesgue-measurable}\} \\ \mathcal{M}_0 &= \{E \in \mathcal{M} : E \subset [a, b]\} \\ \lambda_0 &= \lambda|_{\mathcal{M}_0}.\end{aligned}$$

Form the real Banach space $L^1 = L^1_{\mathbb{R}}([a, b], \mathcal{M}_0, \lambda_0)$ (cf. 6.4.18). Bending the notation, we sometimes abbreviate λ_0 to λ .

6.6.10. Theorem. (Riesz representation theorem²) *With the preceding notations, if $T : L^1 \rightarrow \mathbb{R}$ is a continuous linear form on L^1 , then there exists an essentially bounded measurable function $g : [a, b] \rightarrow \mathbb{R}$ such that*

$$Tf = \int fg \, d\lambda \quad \text{for all } f \in L^1.$$

Proof. We show first that there exists a real number $M \geq 0$ such that

$$(1) \quad |Tu| \leq M\|u\|_1 \quad \text{for all } u \in L^1.$$

Assume to the contrary that no such M exists. Then each positive integer n fails to have the property required of M , so that there exists a $u_n \in L^1$ such that $|Tu_n| > n\|u_n\|_1$. In particular, $Tu_n \neq 0$, therefore $u_n \neq 0$. Writing

$$v_n = \frac{1}{n}\|u_n\|_1^{-1}u_n,$$

we have $\|v_n\|_1 = 1/n$ and

$$|Tv_n| = \frac{1}{n}\|u_n\|_1^{-1}|Tu_n| > \frac{1}{n}\|u_n\|_1^{-1} \cdot n\|u_n\|_1 = 1;$$

thus $v_n \rightarrow 0$ but $Tv_n \not\rightarrow 0 = T0$, contrary to the continuity of T at $0 \in L^1$.

For the rest of the proof, fix a number $M \geq 0$ satisfying (1). {Incidentally, there is a smallest such M , easily seen to be equal to $\sup |Tu|$, where u varies over all elements of L^1 such that $\|u\|_1 \leq 1$; this supremum is called the *norm* of T and is denoted $\|T\|$.} For every $x \in [a, b]$, let

$$\varphi_{[a,x]} : [a, b] \rightarrow \mathbb{R}$$

be the characteristic function of the subinterval $[a, x]$ of $[a, b]$. Then $\varphi_{[a,x]} \in L^1$ and we may apply T to it; define $F : [a, b] \rightarrow \mathbb{R}$ by the formula

$$F(x) = T\varphi_{[a,x]} \quad (x \in [a, b]).$$

² Cf. H.L. Royden, *Real analysis* [Macmillan, 3rd edn., New York, 1988], p. 132.

The function F satisfies the Lipschitz condition

$$(2) \quad |F(x) - F(y)| \leq M|x - y| \quad \text{for all } x, y \in [a, b].$$

For, assuming (say) that $a \leq x \leq y \leq b$, we have

$$\begin{aligned} F(y) - F(x) &= T\dot{\varphi}_{[a,y]} - T\dot{\varphi}_{[a,x]} \\ &= T(\dot{\varphi}_{[a,y]} - \dot{\varphi}_{[a,x]}) \\ &= T(\varphi_{[a,y]} - \varphi_{[a,x]}) \\ &= T\dot{\varphi}_{(x,y)}, \end{aligned}$$

thus

$$\begin{aligned} |F(y) - F(x)| &= |T\dot{\varphi}_{(x,y)}| \\ &\leq M\|\dot{\varphi}_{(x,y)}\|_1 = M \int_a^b \varphi_{(x,y)} d\lambda \\ &= M\lambda((x, y]) = M(y - x) = M|y - x|, \end{aligned}$$

whence the Lipschitz condition (2).

Since F is Lipschitz, it is absolutely continuous (5.1.10, (vi)). By Lebesgue's Fundamental theorem of calculus (5.10.2), there exists a function $g \in \mathcal{L}^1$ such that $F' = g$ a.e. and

$$F(x) = \int_a^x g d\lambda + F(a) \quad \text{for all } x \in [a, b].$$

It follows from the Lipschitz condition (2) that

$$\left| \frac{F(y) - F(x)}{y - x} \right| \leq M$$

for all x, y in $[a, b]$ with $x \neq y$, consequently $|g| \leq M$ a.e., thus $g \in \mathcal{L}^\infty$. Moreover, if $a \leq x \leq y \leq b$ then

$$\begin{aligned} T\dot{\varphi}_{(x,y)} &= F(y) - F(x) = \int_a^y g d\lambda - \int_a^x g d\lambda \\ &= \int [\varphi_{[a,y]} - \varphi_{[a,x]}] g d\lambda = \int \varphi_{(x,y)} g d\lambda; \end{aligned}$$

if I is any of the four possible intervals with endpoints x and y , then $\varphi_{(x,y)} = \varphi_I$ a.e., and we have shown that

$$(3) \quad T\dot{\varphi}_I = \int \varphi_I g d\lambda \quad \text{for every subinterval } I \text{ of } [a, b].$$

For every Lebesgue-measurable set $E \subset [a, b]$, define

$$\nu_1(E) = \int \varphi_E g d\lambda, \quad \nu_2(E) = T\dot{\varphi}_E.$$

Since $\nu_1 = g \cdot \lambda_0$ (cf. 4.7.1), we know that ν_1 is a real measure on \mathcal{M}_0 (cf. 4.7.3).

claim: ν_2 is also a real measure on \mathcal{M}_0 .

Assuming $E = \bigcup_{n=1}^{\infty} E_n$ is the union of a sequence of pairwise disjoint sets in \mathcal{M}_0 , we are to show that

$$\nu_2(E) = \sum_{n=1}^{\infty} \nu_2(E_n).$$

Since $\sum_{n=1}^{\infty} \lambda(E_n) = \lambda(E) \leq b - a < +\infty$, we see that

$$\lambda(E) - \sum_{k=1}^n \lambda(E_k) = \sum_{k=n+1}^{\infty} \lambda(E_k) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

that is, $\int [\varphi_E - \sum_{k=1}^n \varphi_{E_k}] d\lambda \rightarrow 0$; thus

$$\|\varphi_E - \sum_{k=1}^n \varphi_{E_k}\|_1 \rightarrow 0,$$

whence $\sum_{k=1}^n \dot{\varphi}_{E_k} \rightarrow \dot{\varphi}_E$ in L^1 . Since T is linear and continuous,

$$\sum_{k=1}^n T\dot{\varphi}_{E_k} \rightarrow T\dot{\varphi}_E,$$

that is,

$$\nu_2(E) = \lim_{n \rightarrow \infty} \sum_{k=1}^n \nu_2(E_k);$$

in other words $\sum_{n=1}^{\infty} \nu_2(E_n) = \nu_2(E)$, as claimed.

The equation (3) may now be written

$$\nu_2(I) = \nu_1(I) \quad \text{for every subinterval } I \text{ of } [a, b],$$

where ν_2 and ν_1 are both real measures on \mathcal{M}_0 ; it follows from Corollary 6.5.4 that

$$\nu_2(B) = \nu_1(B) \quad \text{for all } B \in \mathcal{B}_0,$$

where \mathcal{B}_0 is the set of all Borel sets $B \subset [a, b]$, that is,

$$(4) \quad T\dot{\varphi}_B = \int \varphi_B g d\lambda \quad \text{for all } B \in \mathcal{B}_0.$$

Indeed, $\nu_2 = \nu_1$ on \mathcal{M}_0 . For, if $E \in \mathcal{M}_0$ then there exists a Borel set B in \mathbb{R} such that $E \subset B$ and $\lambda(B - E) = 0$ (2.4.15); replacing B by $B \cap [a, b]$, we can suppose that $B \in \mathcal{B}_0$. Then $\varphi_E = \varphi_B$ a.e. (with respect to λ_0), thus $\dot{\varphi}_E = \dot{\varphi}_B$ and, citing (4) at the appropriate step,

$$\nu_2(E) = T\dot{\varphi}_E = T\dot{\varphi}_B = \int \varphi_B g d\lambda = \int \varphi_E g d\lambda = \nu_1(E).$$

Thus,

$$(5) \quad T\dot{\varphi}_E = \int \varphi_E g \, d\lambda \quad \text{for all } E \in \mathcal{M}_0.$$

It then follows from linearity that

$$(6) \quad T\dot{f} = \int f g \, d\lambda$$

for every simple function $f : [a, b] \rightarrow \mathbb{R}$ (relative to \mathcal{M}_0), such functions being automatically integrable with respect to the finite measure λ_0 .

Our problem is to show that the equation (6) holds for every $f \in \mathcal{L}^1_{\mathbb{R}}(\lambda_0)$; in view of 4.4.7, we can suppose that $f \geq 0$. If (f_n) is a sequence of simple functions (relative to \mathcal{M}_0) such that $0 \leq f_n \uparrow f$ pointwise on $[a, b]$ (4.1.26), then $\int f_n \, d\lambda_0 \uparrow \int f \, d\lambda_0$ (4.4.3). Since $f - f_n \geq 0$, it follows that $\|f - f_n\|_1 \rightarrow 0$, therefore $T\dot{f}_n \rightarrow T\dot{f}$ by the continuity of T . Also,

$$f_n g \rightarrow f g \quad \text{pointwise on } [a, b],$$

and

$$|f_n g| = f_n |g| \leq f |g| \leq M f \quad \text{a.e.,}$$

therefore

$$\int f_n g \, d\lambda \rightarrow \int f g \, d\lambda$$

by Lebesgue's dominated convergence theorem (4.5.4). We know from (6) that

$$T\dot{f}_n = \int f_n g \, d\lambda \quad \text{for all } n,$$

and passage to the limit yields the desired formula $T\dot{f} = \int f g \, d\lambda$. \diamond

Exercises

1. Give an example where $\|fg\|_{\infty} < \|f\|_{\infty} \|g\|_{\infty}$; that is, the inequality in (5) of Proposition 6.6.5 may be strict.

{Hint: Consider a two-point set $X = \{a, b\}$, the σ -algebra $\mathcal{S} = \mathcal{P}(X) = \{\emptyset, \{a\}, \{b\}, X\}$, and the discrete measure μ on \mathcal{S} (§2.4, Exercise 3).}

2. The proof of Theorem 6.6.9 shows that

$$\sup\{|T\dot{f}| : f \in \mathcal{L}^1_{\mathbb{C}}(\mu), \|f\|_1 \leq 1\} \leq \|g\|_{\infty}.$$

Give an example for which the inequality is strict.

{Hint: With (X, \mathcal{S}) the measurable space in the hint for Exercise 1, consider the measure μ given by

$$\mu(\emptyset) = 0, \quad \mu(\{a\}) = 1, \quad \mu(\{b\}) = +\infty, \quad \mu(X) = +\infty.$$

Then $\mathcal{L}_\mathbb{C}^1$ consists of the scalar multiples of $\varphi_{\{a\}}$, \emptyset is the only set of measure zero, and the linear forms on the one-dimensional vector space $L_\mathbb{C}^1 = \mathcal{L}_\mathbb{C}^1$ are the scalar multiples of the linear form $f \mapsto \int f d\mu$ (in other words, the linear form $c\varphi_{\{a\}} \mapsto c$). Let $g = \varphi_{\{b\}}$ and contemplate $Tf = \int fg d\mu$ ($f \in \mathcal{L}_\mathbb{C}^1$).}

3. With notations as in Definition 6.6.6, $\|uv\|_\infty \leq \|u\|_\infty \|v\|_\infty$ for all $u, v \in L_\mathbb{C}^\infty$; so to speak, the norm on the algebra $L_\mathbb{C}^\infty$ is *submultiplicative*. {A Banach space with an associative and distributive multiplication that commutes with scalar multiplication (cf. §5.1, Exercise 4) and for which the norm is submultiplicative is called a *Banach algebra*.}

4. (i) The function g of the Riesz representation theorem (6.6.10) is a.e. unique.

(ii) Extend the Riesz theorem to Lebesgue measure on \mathbb{R} , that is, to continuous linear forms on $L_\mathbb{R}^1(\mathbb{R}, \mathcal{M}, \lambda)$.

{Hint: (i) Cf. 4.4.24. (ii) For each positive integer n , regard $L^1[-n, n]$ as a linear subspace of $L^1(\mathbb{R})$ by extending functions $f: [-n, n] \rightarrow \mathbb{R}$ to be 0 on $\mathbb{R} - [-n, n]$. If T is a continuous linear form on $L^1(\mathbb{R})$, apply Theorem 6.6.10 to each of the restrictions $T_n = T|_{L^1[-n, n]}$ to produce an essentially bounded function $g_n: [-n, n] \rightarrow \mathbb{R}$, then argue that there is an essentially bounded function $g: \mathbb{R} \rightarrow \mathbb{R}$ such that $g|_{[-n, n]} = g_n$ a.e. for every n .}

5. Extend the Riesz theorem (6.6.10) to the complex case, that is, to continuous \mathbb{C} -linear forms $T: L_\mathbb{C}^1(\lambda_0) \rightarrow \mathbb{C}$.

{Hint: Regard $L_\mathbb{R}^1 \subset L_\mathbb{C}^1$ in the obvious way (remarks following Definition 6.4.4) and apply 6.6.10 to the \mathbb{R} -linear forms $R, S: L_\mathbb{R}^1 \rightarrow \mathbb{R}$ defined by $Rf = \operatorname{Re}(Tf)$, $Sf = \operatorname{Im}(Tf)$ for all $f \in \mathcal{L}_\mathbb{R}^1(\lambda_0)$.}

6. Let $g \in \mathcal{L}_\mathbb{C}^\infty(\mu)$ and let $T: L_\mathbb{C}^1(\mu) \rightarrow \mathbb{C}$ be the linear form defined by $Tf = \int fg d\mu$. We know from Theorem 6.6.9 that T is continuous; writing

$$\|T\| = \sup\{|Tf| : f \in \mathcal{L}_\mathbb{C}^1, \|f\|_1 \leq 1\},$$

we see from the inequality $\|fg\|_1 \leq \|f\|_1 \|g\|_\infty$ that $\|T\| \leq \|g\|_\infty$. It can happen that $\|T\| < \|g\|_\infty$ (Exercise 2).

Prove: If μ is σ -finite then $\|T\| = \|g\|_\infty$.

{Hint: Write $g = u|g|$ with u measurable and $|u| = 1$ (§6.4, Exercise 3), and let $S: L_\mathbb{C}^1 \rightarrow \mathbb{C}$ be the linear form $Sf = \int f|g| d\mu$. From $|g| = \bar{u}g$ and $\|fu\|_1 = \|f\|_1 = \|f\bar{u}\|_1$ for all $f \in \mathcal{L}_\mathbb{C}^1$, conclude that $\|T\| = \|S\|$; we can suppose, therefore, that $g \geq 0$. Given any $\epsilon > 0$, it will suffice to show that $\|T\| \geq \|g\|_\infty - \epsilon$. If $E = \{x: g(x) > \|g\|_\infty - \epsilon\}$, necessarily $\mu(E) > 0$ (the alternative is that $0 \leq g \leq \|g\|_\infty - \epsilon$ a.e., whence the absurdity $\|g\|_\infty \leq \|g\|_\infty - \epsilon$). Let $F \in \mathcal{S}$ with $F \subset E$ and $0 < \mu(F) < +\infty$; then $f = (1/\mu(F))\varphi_F$ is integrable, $\|f\|_1 = 1$, and integration of the inequality $fg \geq (\|g\|_\infty - \epsilon)f$ leads to paydirt.}

7. (i) With notations as in Exercise 6, if μ is σ -finite then g is 'essentially uniquely' determined by T , in the sense that if also $h \in \mathcal{L}_{\mathbb{C}}^{\infty}(\mu)$ and $Tf = \int fg d\mu = \int fh d\mu$ for all $f \in \mathcal{L}_{\mathbb{C}}^1(\mu)$, then $g = h$ a.e.

{Hint: Part (ii) of Theorem 4.7.2.}

(ii) Give an example of a measure μ (not σ -finite) for which 'essential uniqueness' in the foregoing sense fails.

{Hint: Exercise 2.}

6.7. L^p ($1 < p < +\infty$)

Notations fixed throughout the section: (X, \mathcal{S}, μ) is a measure space, p a real number, $1 < p < +\infty$, and $q = \frac{p}{p-1}$ (the positive real number 'conjugate' to p); thus $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, in other words, $p + q = pq$.

6.7.1. *Definition.* A complex-valued function $f : X \rightarrow \mathbb{C}$ is said to be **p -th power integrable** if

(i) f is measurable (with respect to \mathcal{S}), and

(ii) $|f|^p$ is integrable (with respect to μ), that is, $|f|^p \in \mathcal{L}^1$.

The set of all such functions f is denoted

$$\mathcal{L}_{\mathbb{C}}^p(X, \mathcal{S}, \mu),$$

briefly $\mathcal{L}_{\mathbb{C}}^p$, and, for such a function f , we write

$$\|f\|_p = \left(\int |f|^p d\mu \right)^{1/p},$$

called the **\mathcal{L}^p -norm** (or **p -norm**) of f .

The foregoing is a slight abuse of the term "norm"; as we shall see below (in 6.7.4), $\mathcal{L}_{\mathbb{C}}^p$ is a complex vector space for the pointwise linear operations, and $f \mapsto \|f\|_p$ is in fact a *seminorm* on $\mathcal{L}_{\mathbb{C}}^p$. The proof of the triangle inequality for this seminorm is based on the extension to integrals of an inequality proved in Chapter 3 (3.1.12):

6.7.2. Theorem. (Hölder's inequality) If $f \in \mathcal{L}_{\mathbb{C}}^p$ and $g \in \mathcal{L}_{\mathbb{C}}^q$ ($q = \frac{p}{p-1}$), then $fg \in \mathcal{L}_{\mathbb{C}}^1$ and

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Proof. At any rate, fg is measurable (6.4.2); we have to prove that it is integrable. Let $\alpha = \|f\|_p$, $\beta = \|g\|_q$.

If $\alpha = 0$ or $\beta = 0$, then $f = 0$ a.e. or $g = 0$ a.e., fg ($= 0$ a.e.) is integrable, and the asserted inequality is obvious (6.4.12).

Suppose $\alpha > 0$ and $\beta > 0$. For each $x \in X$, application of Proposition 3.1.3 (with $a = |f(x)|/\alpha$ and $b = |g(x)|/\beta$) yields the inequality

$$\frac{|f(x)g(x)|}{\alpha\beta} \leq \frac{1}{p} \cdot \frac{|f(x)|^p}{\alpha^p} + \frac{1}{q} \cdot \frac{|g(x)|^q}{\beta^q},$$

thus the functions f and g satisfy (identically) the inequality

$$(*) \quad \frac{1}{\alpha\beta} |fg| \leq \frac{1}{p} \cdot \frac{|f|^p}{\alpha^p} + \frac{1}{q} \cdot \frac{|g|^q}{\beta^q}.$$

The sum on the right side of $(*)$ is integrable by hypothesis, therefore so is fg (cf. 4.4.20 and 6.4.6). Since

$$\int |f|^p d\mu = \alpha^p \quad \text{and} \quad \int |g|^q d\mu = \beta^q$$

(by the definition of α and β), integration of $(*)$ yields

$$\frac{1}{\alpha\beta} \|fg\|_1 \leq \frac{1}{p} + \frac{1}{q} = 1,$$

thus $\|fg\|_1 \leq \alpha\beta$ as claimed. \diamond

6.7.3. Theorem. (Minkowski's inequality) If $f, g \in \mathcal{L}_{\mathbb{C}}^p$ then $f + g \in \mathcal{L}_{\mathbb{C}}^p$ and

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Proof. We know that $f + g$ is measurable (6.4.2) and we have to show that $|f + g|^p$ is integrable. Writing $k = |f| \cup |g|$, we have

$$|f + g| \leq |f| + |g| \leq 2k$$

and, since the function $t \mapsto t^p$ is an order isomorphism $[0, +\infty) \rightarrow [0, +\infty)$,

$$k^p = |f|^p \cup |g|^p \leq |f|^p + |g|^p,$$

therefore

$$|f + g|^p \leq 2^p k^p \leq 2^p (|f|^p + |g|^p) \in \mathcal{L}^1;$$

it follows that $|f + g|^p$ is integrable (4.4.5), thus $f + g \in \mathcal{L}_{\mathbb{C}}^p$. The integrability of the function

$$|f + g|^p = |f + g|^{pq-q} = (|f + g|^{p-1})^q$$

shows that $|f + g|^{p-1} \in \mathcal{L}^q$; writing $h = |f + g|^{p-1}$, we have $h^q = |f + g|^p \in \mathcal{L}^1$ and

$$(1) \quad \|h\|_q = \left(\int h^q d\mu \right)^{1/q} = \left(\int |f + g|^p d\mu \right)^{1/q} = (\|f + g\|_p)^{p/q}.$$

Also

$$(2) \quad |f + g|^p = |f + g|h \leq |f|h + |g|h = |fh| + |gh|;$$

by the preceding theorem, fh and gh are integrable, and, citing Hölder's inequality at the appropriate step, integration of (2) yields

$$\int |f + g|^p d\mu \leq \int |fh| d\mu + \int |gh| d\mu \leq \|f\|_p \|h\|_q + \|g\|_p \|h\|_q,$$

that is, citing (1),

$$(\|f + g\|_p)^p \leq (\|f\|_p + \|g\|_p) \|h\|_q = (\|f\|_p + \|g\|_p) (\|f + g\|_p)^{p/q};$$

it follows that

$$(\|f + g\|_p)^{p-p/q} \leq \|f\|_p + \|g\|_p$$

(even if $\|f + g\|_p = 0$) and the observation that $p - p/q = 1$ completes the proof. \diamond

For the case $p = 2$, item (iii) of the following theorem is known as the *Riesz-Fischer theorem*.¹

6.7.4. Theorem. (i) $\mathcal{L}_{\mathbb{C}}^p$ is a complex vector space for the pointwise linear operations, and $f \mapsto \|f\|_p$ is a seminorm on $\mathcal{L}_{\mathbb{C}}^p$.

(ii) The set $\mathcal{N} = \mathcal{N}(X, \mathcal{S}, \mu)$ of measurable complex functions f on X such that $f = 0$ a.e. is a linear subspace of $\mathcal{L}_{\mathbb{C}}^p$, and

$$\mathcal{N} = \{f \in \mathcal{L}_{\mathbb{C}}^p : \|f\|_p = 0\}.$$

(iii) If (f_n) is a sequence in $\mathcal{L}_{\mathbb{C}}^p$ with

$$\|f_m - f_n\|_p \rightarrow 0 \quad \text{as } m, n \rightarrow \infty,$$

then there exists a function $f \in \mathcal{L}_{\mathbb{C}}^p$ such that $\|f_n - f\|_p \rightarrow 0$; moreover, any two such functions f are equal a.e.

Proof. (i) If $f \in \mathcal{L}_{\mathbb{C}}^p$ and $c \in \mathbb{C}$, then cf is measurable (6.4.2) and $|cf|^p = |c|^p |f|^p$ is integrable, therefore $cf \in \mathcal{L}_{\mathbb{C}}^p$ and $\|cf\|_p = |c| \|f\|_p$; the preceding theorem then completes the proof of (i).

(ii) If $f \in \mathcal{N}$ then f is measurable and $|f|^p = 0$ a.e., therefore $f \in \mathcal{L}_{\mathbb{C}}^p$ and $\int |f|^p d\mu = 0$, thus $\|f\|_p = 0$. Conversely, if $f \in \mathcal{L}_{\mathbb{C}}^p$ and $\|f\|_p = 0$, then $\int |f|^p d\mu = 0$, whence $|f|^p = 0$ a.e. (4.4.21), therefore $f \in \mathcal{N}$.

(iii) The proof is similar to that of the analogous property of $\mathcal{L}_{\mathbb{C}}^1$ (cf. 6.4.13). Let (f_n) be a sequence in $\mathcal{L}_{\mathbb{C}}^p$ such that $\|f_m - f_n\|_p \rightarrow 0$ as $m, n \rightarrow \infty$. To simplify the notations, let us abbreviate $\|f\|_p$ to $\|f\|$, for $f \in \mathcal{L}_{\mathbb{C}}^p$. By the triangle inequality (cf. 6.4.10) it clearly suffices to find an $f \in \mathcal{L}_{\mathbb{C}}^p$ such that $\|f_{n_k} - f\| \rightarrow 0$ for some subsequence (f_{n_k}) of (f_n) .

¹ F. Riesz (1880–1956) and E. Fischer (1875–1954).

(cf. the proof of 6.1.26). Thus, passing to a subsequence, we can suppose that

$$\|f_{n+1} - f_n\| \leq 2^{-n} \quad \text{for all } n.$$

Write $\alpha = \sum_{n=1}^{\infty} \|f_{n+1} - f_n\| \leq \sum_{n=1}^{\infty} 2^{-n} = 1$. Let $f_0 = 0$ and define

$$g_n = \sum_{k=1}^n |f_k - f_{k-1}| \quad \text{for } n = 1, 2, 3, \dots;$$

it follows from (i) that $g_n \in \mathcal{L}^p$, therefore $(g_n)^p \in \mathcal{L}^1$ for all n . Clearly $0 \leq g_n \uparrow$, therefore also $0 \leq (g_n)^p \uparrow$; by Minkowski's inequality,

$$\|g_n\| \leq \sum_{k=1}^n \|f_k - f_{k-1}\| = \|f_1\| + \sum_{k=2}^n \|f_k - f_{k-1}\| \leq \|f_1\| + \alpha,$$

thus

$$\int (g_n)^p d\mu = \|g_n\|^p \leq (\|f_1\| + \alpha)^p < +\infty$$

for all n ; by the monotone convergence theorem, there exists an $h \in \mathcal{L}^1$ such that $(g_n)^p \uparrow h$ a.e. Redefining the f_n and h to be zero on a suitable null set (i.e., on a suitable measurable set of measure 0), we can suppose that

$$(*) \quad 0 \leq (g_n)^p \uparrow h \quad \text{pointwise on } X.$$

Let $g = h^{1/p}$; then g is measurable and $g^p = h \in \mathcal{L}^1$, thus $g \in \mathcal{L}^p$. Also,

$$0 \leq g_n \uparrow g$$

by (*). For each $x \in X$,

$$\sum_{k=1}^n |f_k(x) - f_{k-1}(x)| = g_n(x) \uparrow g(x) < +\infty,$$

therefore the series $\sum_{k=1}^{\infty} [f_k(x) - f_{k-1}(x)]$ is (absolutely) convergent and we may define $f(x)$ to be its sum:

$$f(x) = \lim_{n \rightarrow \infty} \sum_{k=1}^n [f_k(x) - f_{k-1}(x)] = \lim_{n \rightarrow \infty} f_n(x).$$

Thus $f_n \rightarrow f$ pointwise, therefore f is measurable (6.4.3). Also, for every

$x \in X$,

$$\begin{aligned} |f_n(x)| &= \left| \sum_{k=1}^n [f_k(x) - f_{k-1}(x)] \right| \\ &\leq \sum_{k=1}^n |f_k(x) - f_{k-1}(x)| = g_n(x) \leq g(x) \end{aligned}$$

for all n ; passage to the limit yields $|f| \leq g$, therefore $|f|^p \leq g^p = h \in \mathcal{L}^1$, whence $|f|^p$ is integrable and so $f \in \mathcal{L}_\mathbb{C}^p$.

Next, we show that $\|f_n - f\| \rightarrow 0$; the proof will make use of Fatou's lemma (4.5.5). At any rate, by the preceding paragraph, $f_n - f \in \mathcal{L}_\mathbb{C}^p$ for all n , so $\|f_n - f\|$ makes sense. Let $\epsilon > 0$. Choose an index N such that

$$m, n \geq N \Rightarrow \|f_m - f_n\| \leq \epsilon.$$

Fix an index $m \geq N$. Then

$$(**) \quad \int |f_m - f_n|^p d\mu \leq \epsilon^p \quad \text{for all } n \geq N;$$

also, as $n \rightarrow \infty$,

$$|f_m - f_n|^p \rightarrow |f_m - f|^p \in \mathcal{L}^1,$$

thus

$$\liminf_n |f_m - f_n|^p = \lim_{n \rightarrow \infty} |f_m - f_n|^p = |f_m - f|^p \in \mathcal{L}^1,$$

and Fatou's lemma yields, in view of (**),

$$\int |f_m - f|^p d\mu \leq \liminf_n \int |f_m - f_n|^p d\mu \leq \epsilon^p,$$

whence $\|f_m - f\| \leq \epsilon$ (for every $m \geq N$).

Finally, if also $f_* \in \mathcal{L}_\mathbb{C}^p$ with $\|f_n - f_*\| \rightarrow 0$, then

$$\|f - f_*\| \leq \|f - f_n\| + \|f_n - f_*\|$$

for all n ; passage to the limit yields $\|f - f_*\| = 0$, whence $f - f_* = 0$ a.e. by (i). \diamond

6.7.5. *Definition.* With notations as in 6.7.4, the quotient vector space $\mathcal{L}_\mathbb{C}^p/\mathcal{N}$ is denoted

$$L_\mathbb{C}^p(X, \mathcal{S}, \mu),$$

briefly $L_\mathbb{C}^p$. For $u \in L_\mathbb{C}^p$, say $u = \dot{f} = f + \mathcal{N}$, where $f \in \mathcal{L}_\mathbb{C}^p$, one writes

$$\|u\|_p = \|f\|_p,$$

called the *norm* (or L^p -*norm*) of u ; if $f, g \in \mathcal{L}_\mathbb{C}^p$ and $\dot{f} = \dot{g}$, then $f - g \in \mathcal{N}$, $f = g$ a.e., and $\|f\|_p = \|g\|_p$, thus $\|u\|_p$ depends only on the coset u , not on the particular function $f \in \mathcal{L}_\mathbb{C}^p$ selected to represent it.

6.7.6. Corollary. *With notations as in the preceding definition, $L_{\mathbb{C}}^p$ is a (complex) Banach space with $u \mapsto \|u\|_p$ as norm.*

Proof. The proof is similar to that for $L_{\mathbb{C}}^1$ (6.4.18), with the requisite completeness supplied by Theorem 6.7.4. \diamond

A consequence of Hölder's inequality is that each function in $\mathcal{L}_{\mathbb{C}}^q$ induces a linear form on $\mathcal{L}_{\mathbb{C}}^p$ (and, ultimately, on the Banach space $L_{\mathbb{C}}^p$), continuous in an appropriate sense:

6.7.7. Theorem. *Let $g \in \mathcal{L}_{\mathbb{C}}^q$ ($q = \frac{p}{p-1}$).*

(i) *The formula*

$$L(f) = \int fg d\mu \quad (f \in \mathcal{L}_{\mathbb{C}}^p)$$

defines a linear form L on $\mathcal{L}_{\mathbb{C}}^p$, such that

$$|L(f)| \leq \|f\|_p \|g\|_q \quad \text{for all } f \in \mathcal{L}_{\mathbb{C}}^p.$$

(ii) *L is continuous in the sense that*

$$\|f_n - f\|_p \rightarrow 0 \quad \Rightarrow \quad L(f_n) \rightarrow L(f).$$

(iii) *Moreover,*

$$\|g\|_q = \sup\{|L(f)| : f \in \mathcal{L}_{\mathbb{C}}^p, \|f\|_p \leq 1\}.$$

Proof. (i) The indicated integrals exist by Theorem 6.7.2 and, for $f \in \mathcal{L}_{\mathbb{C}}^p$,

$$|L(f)| = \left| \int fg d\mu \right| \leq \int |fg| d\mu = \|fg\|_1 \leq \|f\|_p \|g\|_q$$

by Hölder's inequality. The linearity of L follows from the linearity of integration (6.4.5).

(ii) Immediate from (i).

(iii) The asserted equality is obvious if $g = 0$ a.e. (both sides are 0).

Suppose $\|g\|_q > 0$. If $f \in \mathcal{L}_{\mathbb{C}}^p$ and $\|f\|_p \leq 1$, then $|L(f)| \leq \|g\|_q$ by (i), so the indicated supremum is finite and, writing M for this supremum, we have $M \leq \|g\|_q$; the problem is to prove the reverse inequality. Better yet, we shall show that there exists a function $f \in \mathcal{L}_{\mathbb{C}}^p$ such that $\|f\|_p = 1$ and $L(f) = \|g\|_q$, which will imply that $\|g\|_q = |L(f)| \leq M$.

Write $g = u|g|$ with u a measurable function such that $|u| = 1$ (§6.4, Exercise 3) and define $h = \bar{u}|g|^{q-1}$; a suitable scalar multiple of h will yield the desired function f . At any rate, h is measurable and

$$|h|^p = (|g|^{q-1})^p = |g|^{pq-p} = |g|^q \in \mathcal{L}^1,$$

therefore $h \in \mathcal{L}_{\mathbb{C}}^p$; moreover,

$$(\|h\|_p)^p = \int |h|^p d\mu = \int |g|^q d\mu = (\|g\|_q)^q,$$

thus

$$(1) \quad \|h\|_p = (\|g\|_q)^{q/p} = (\|g\|_q)^{q-1}.$$

On the other hand,

$$\begin{aligned} L(h) &= \int hg d\mu = \int (\bar{u} |g|^{q-1}) g d\mu \\ (2) \quad &= \int |g|^{q-1} (\bar{u} g) d\mu = \int |g|^{q-1} |g| d\mu \\ &= \int |g|^q d\mu = (\|g\|_q)^q \\ &= \|g\|_q (\|g\|_q)^{q-1}. \end{aligned}$$

Since $h \in \mathcal{L}_\mathbb{C}^p$, its scalar multiple $f = (\|g\|_q)^{1-q} h$ also belongs to $\mathcal{L}_\mathbb{C}^p$, $\|f\|_p = 1$ by (1), and $L(f) = \|g\|_q$ by (2). \diamond

6.7.8. It is true, conversely, that if $L : \mathcal{L}_\mathbb{C}^p(\mu) \rightarrow \mathbb{C}$ is a linear form that is continuous in the sense of (ii) of the preceding theorem, then there exists a function $g \in \mathcal{L}_\mathbb{C}^q$ such that $L(f) = \int f g d\mu$ for all $f \in \mathcal{L}_\mathbb{C}^p$. The general case can be inferred from the case that the measure μ is finite;² we conclude this section with the proof for that special case.

These results are not cited elsewhere in the text and can be omitted, but they can provide the reader with a toe-hold on an important subject (duality of L^p -spaces). The special case considered here is accessible enough to be derived from earlier results in this section, yet too complicated to be parceled out as an “exercise” (in name only). The exposition is based on that in the book of H.L. Royden,³ where the reader will find more general versions of the results presented here.

6.7.9. Lemma. *The following conditions on a linear form $L : \mathcal{L}_\mathbb{C}^p \rightarrow \mathbb{C}$ are equivalent:*

- (a) $\|f_n - f\|_p \rightarrow 0 \Rightarrow L(f_n) \rightarrow L(f)$;
- (b) *there exists a constant $M \geq 0$ such that $|L(f)| \leq M\|f\|_p$ for all $f \in \mathcal{L}_\mathbb{C}^p$;*
- (c) *the set of complex numbers $\{L(f) : f \in \mathcal{L}_\mathbb{C}^p, \|f\|_p \leq 1\}$ is bounded.*

When the foregoing conditions are verified, the correspondence $f \mapsto L(f)$ defines a continuous linear form on the Banach space $L_\mathbb{C}^p = \mathcal{L}_\mathbb{C}^p / \mathcal{N}$ of 6.7.6.

Proof. (a) \Rightarrow (b): Note first that $\|f\|_p = 0 \Rightarrow L(f) = 0$ (consider the sequence $f_n = 0$ for all n). Assume to the contrary that no such M exists. Then, for every positive integer n , there exists a function $g_n \in \mathcal{L}_\mathbb{C}^p$

² Cf. H.L. Royden, *Real analysis* [3rd. edn., Macmillan, New York, 1988], p. 286, Theorem 30.

³ *Op. cit.*, Chapter 11, §7 (pp. 282-287).

such that $|L(g_n)| > n\|g_n\|_p$ (in particular, $\|g_n\|_p > 0$ by the preceding remark); the functions $f_n = (n\|g_n\|_p)^{-1}g_n$ then satisfy

$$\|f_n\|_p = 1/n \quad \text{and} \quad |L(f_n)| = (n\|g_n\|_p)^{-1}|L(g_n)| > 1,$$

thus $\|f_n\|_p \rightarrow 0$ but $L(f_n) \not\rightarrow 0$, contrary to (a).

(b) \Rightarrow (c): With M as in (b), $|L(f)| \leq M$ whenever $\|f\|_p \leq 1$.

(c) \Rightarrow (b): Let M be an upper bound for the numbers $|L(f)|$ ($f \in \mathcal{L}_{\mathbb{C}}^p$, $\|f\|_p \leq 1$). Given any $g \in \mathcal{L}_{\mathbb{C}}^p$, we assert that $|L(g)| \leq M\|g\|_p$. If $\|g\|_p = 0$ then, for every positive integer n , $\|ng\|_p = n\|g\|_p = 0 < 1$, therefore $|L(ng)| \leq M$ by hypothesis; the validity of $|L(g)| \leq M/n$ for all n means that $L(g) = 0$, thus the desired inequality holds trivially. On the other hand, if $\|g\|_p > 0$ then the function $f = (\|g\|_p)^{-1}g$ satisfies $\|f\|_p = 1$, therefore $|L(f)| \leq M$, whence $|L(g)| \leq M\|g\|_p$ by the linearity of L .

(b) \Rightarrow (a): $|L(f_n) - L(f)| = |L(f_n - f)| \leq M\|f_n - f\|_p$.

Finally, as noted in the proof of (a) \Rightarrow (b), such a linear form L satisfies $\|f\|_p = 0 \Rightarrow L(f) = 0$; it follows that if $u \in L_{\mathbb{C}}^1 = \mathcal{L}_{\mathbb{C}}^1/\mathcal{N}$, say $u = \dot{f} = f + \mathcal{N}$, then the number $L(f)$ depends only on u and not on the particular function f chosen from the coset. Thus, the correspondence

$$u \mapsto L(f) \quad (u = \dot{f} \in L_{\mathbb{C}}^1)$$

is well-defined, it is clearly a linear form on $L_{\mathbb{C}}^1$, and it follows from the condition (b) that this linear form is continuous for the metric topology on the Banach space $L_{\mathbb{C}}^1$ derived from its norm (see the remark following 6.4.15). \diamond

6.7.10. Lemma. Suppose the measure space (X, \mathcal{S}, μ) is finite. If $g: X \rightarrow \mathbb{C}$ is a μ -integrable function such that the set of complex numbers

$$\left\{ \int fg d\mu : f \text{ simple}, \|f\|_p \leq 1 \right\}$$

is bounded, then $g \in \mathcal{L}_{\mathbb{C}}^q$.

Proof. Recall that $1 < p < +\infty$ and $1/p + 1/q = 1$. Since μ is finite it is clear that, for every simple function f , $f \in \mathcal{L}_{\mathbb{C}}^p$ and $fg \in \mathcal{L}_{\mathbb{C}}^1$. Let

$$M = \sup \left\{ \left| \int fg d\mu \right| : f \text{ simple}, \|f\|_p \leq 1 \right\};$$

by the arguments in 6.7.9, it is clear that

$$(*) \quad \left| \int fg d\mu \right| \leq M\|f\|_p$$

for every simple function f .

Consider first the case that g is real-valued; we will prove that $g \in \mathcal{L}_{\mathbb{R}}^q$ assuming only that $(*)$ holds for all real-valued simple functions f . Let $E = \{x: g(x) \geq 0\}$ (a measurable set) and define $u = \varphi_E - \varphi_{X-E}$; then

u is a simple function, $|u| = 1$ on X , and $g = u|g|$. Let (f_n) be a sequence of simple functions such that

$$0 \leq f_n \uparrow |g|^q;$$

to prove that $g \in \mathcal{L}_{\mathbb{R}}^q$, we must show that $|g|^q \in \mathcal{L}_{\mathbb{R}}^1$, and for this it suffices to show that the sequence $\int f_n d\mu$ is bounded above (4.5.1). Let

$$h_n = u f_n^{1/p} \quad (n = 1, 2, 3, \dots);$$

since u and the f_n are real-valued simple functions, so are the h_n , therefore

$$\int h_n g d\mu \leq M \|h_n\|_p$$

by the hypothesis on g . Since $|g| \geq f_n^{1/q}$, we have

$$h_n g = u f_n^{1/p} \cdot g = f_n^{1/p} \cdot u g = f_n^{1/p} \cdot |g| \geq f_n^{1/p} \cdot f_n^{1/q} = f_n^{1/p+1/q} = f_n,$$

therefore

$$(i) \quad \int f_n d\mu \leq \int h_n g d\mu \leq M \|h_n\|_p;$$

but

$$(ii) \quad \|h_n\|_p = \left(\int |h_n|^p d\mu \right)^{1/p} = \left(\int f_n d\mu \right)^{1/p},$$

and it follows from (i) and (ii) that

$$\int f_n d\mu \leq M \left(\int f_n d\mu \right)^{1/p},$$

whence $\left(\int f_n d\mu \right)^{1-1/p} \leq M$ (even if $\int f_n d\mu = 0$), that is, $\left(\int f_n d\mu \right)^{1/q} \leq M$. Thus,

$$\int f_n d\mu \leq M^q \quad \text{for all } n,$$

which completes the proof that $g \in \mathcal{L}_{\mathbb{R}}^q$ (when g is real-valued).

Suppose now that g is complex-valued, and write $g = g_1 + i g_2$ with $g_1, g_2 \in \mathcal{L}_{\mathbb{R}}^1$. For every real-valued simple function f ,

$$\int f g d\mu = \int f g_1 d\mu + i \int f g_2 d\mu$$

is the decomposition of the left member into its real and imaginary parts; since, by hypothesis, the left member remains bounded as f varies over all such functions satisfying $\|f\|_p \leq 1$, the same is true of its real and imaginary parts, consequently $g_1, g_2 \in \mathcal{L}_{\mathbb{R}}^q$ by the first part of the proof, whence $g \in \mathcal{L}_{\mathbb{C}}^q$ by Theorem 6.7.3. \diamond

6.7.11. Theorem. (Riesz representation theorem) Let (X, \mathcal{S}, μ) be a finite measure space, let $1 < p < +\infty$, and let $L : \mathcal{L}_{\mathbb{C}}^p \rightarrow \mathbb{C}$ be a linear form satisfying the continuity conditions of 6.7.9. Then there exists a function $g \in \mathcal{L}_{\mathbb{C}}^q$ ($q = \frac{p}{p-1}$) such that

$$L(f) = \int f g d\mu \quad \text{for all } f \in \mathcal{L}_{\mathbb{C}}^p.$$

Moreover,

$$\|g\|_q = \sup\{|L(f)| : f \in \mathcal{L}_{\mathbb{C}}^p, \|f\|_p \leq 1\},$$

and any two such functions g are equal to each other almost everywhere.

Proof. Define a set function $\nu : \mathcal{S} \rightarrow \mathbb{C}$ by the formula

$$\nu(E) = L(\varphi_E) \quad (E \in \mathcal{S});$$

since $\varphi_{E \cup F} = \varphi_E + \varphi_F$ when $E \cap F = \emptyset$, it is clear from the linearity of L that ν is finitely additive. In fact, ν is countably additive. For, suppose $E = \bigcup_{k=1}^{\infty} E_k$, where (E_k) is a sequence of pairwise disjoint measurable sets; we are to show that the series $\sum_{k=1}^{\infty} \nu(E_k)$ is convergent with sum $\nu(E)$. Writing $F_n = \bigcup_{k=1}^n E_k$, we have $\nu(F_n) = \sum_{k=1}^n \nu(E_k)$ by finite additivity, so the problem is to show that $\nu(F_n) \rightarrow \nu(E)$. Since $F_n \uparrow E$, we have $E - F_n \downarrow \emptyset$, therefore $\mu(E - F_n) \downarrow 0$ by the finiteness of μ (2.6.3); then

$$\begin{aligned} \|\varphi_E - \varphi_{F_n}\|_p &= \|\varphi_{E-F_n}\|_p = \left(\int |\varphi_{E-F_n}|^p d\mu \right)^{1/p} \\ &= \left(\int \varphi_{E-F_n} d\mu \right)^{1/p} = (\mu(E - F_n))^{1/p} \downarrow 0, \end{aligned}$$

thus $\|\varphi_E - \varphi_{F_n}\|_p \rightarrow 0$, therefore

$$\nu(F_n) = L(\varphi_{F_n}) \rightarrow L(\varphi_E) = \nu(E)$$

by the continuity assumption on L .

Summarizing, ν is a complex measure (6.5.1) on the finite measure space (X, \mathcal{S}, μ) . Moreover, ν is absolutely continuous with respect to μ ; for, if $\mu(E) = 0$ then $\|\varphi_E\|_p = 0$, therefore $\nu(E) = L(\varphi_E) = 0$, as shown at the beginning of the proof of 6.7.9. By the Radon-Nikodym theorem (cf. §6.5, Exercise 2), there exists a μ -integrable function $g \in \mathcal{L}_{\mathbb{C}}^1$ such that

$$\int \varphi_E g d\mu = \nu(E) = L(\varphi_E) \quad \text{for all } E \in \mathcal{S}.$$

It then follows from linearity that

$$(*) \quad \int f g d\mu = L(f)$$

for all simple functions f ; in view of the continuity condition on L , it follows from the preceding lemma (6.7.10) that $g \in \mathcal{L}_{\mathbb{C}}^q$.

It remains to show that the formula (*) is verified for every $f \in \mathcal{L}_{\mathbb{C}}^p$. By linearity, it suffices to consider $f \geq 0$ (first write $f = f_1 + if_2$ with f_1 and f_2 real-valued, then write f_1, f_2 as the difference of integrable functions ≥ 0). Choose a sequence (f_n) of simple functions such that $0 \leq f_n \uparrow f$. Then $f - f_n \downarrow 0$, where $f - f_n \in \mathcal{L}_{\mathbb{R}}^p$ (6.7.3), thus

$$|f - f_n|^p \in \mathcal{L}^1 \quad \text{and} \quad |f - f_n|^p \downarrow 0,$$

consequently $\int |f - f_n|^p d\mu \downarrow 0$ by the monotone convergence theorem (4.5.3), whence $\|f - f_n\|_p \rightarrow 0$. By the continuity condition on L , $L(f_n) \rightarrow L(f)$; also, by Hölder's inequality (6.7.2),

$$\|f_n g - f g\|_1 \leq \|f_n - f\|_p \|g\|_q \rightarrow 0,$$

therefore $\int f_n g d\mu \rightarrow \int f g d\mu$ (6.4.6). Since

$$L(f_n) = \int f_n g d\mu \quad \text{for all } n$$

by (*), passage to the limit yields the desired representation $L(f) = \int f g d\mu$.

The formula for $\|g\|_q$ is given in (iii) of 6.7.7. Finally, if $g_1, g_2 \in \mathcal{L}_{\mathbb{C}}^q$ are two functions inducing the same linear form L , then $g_1 - g_2$ induces the zero linear form, therefore $\|g_1 - g_2\|_q = 0$ by the formula just mentioned, in other words, $g_1 - g_2 = 0$ a.e. \diamond

Exercises

1. The boundedness hypothesis in Lemma 6.7.10 follows from the weaker hypothesis that the set

$$\left\{ \int f g d\mu : f \text{ simple and real-valued, } \|f\|_p \leq 1 \right\}$$

is bounded.

{Hint: Write $f = f_1 + if_2$ with f_1, f_2 real-valued.}

2. Let E be a normed space (over \mathbb{R} or \mathbb{C}) and let f be a linear form on E . The following conditions on f are equivalent:

(a) f is continuous for the norm topology on E derived from the norm-metric $(x, y) \mapsto \|x - y\|$ on E (cf. 6.4.10, 3.3.2), in other words, for sequences (x_n) in E ,

$$\|x_n - x\| \rightarrow 0 \Rightarrow f(x_n) \rightarrow f(x);$$

(b) f is continuous at $0 \in E$;

(c) the set $\{|f(x)| : x \in E, \|x\| \leq 1\}$ is bounded;

(d) there exists a constant $M \geq 0$ such that $|f(x)| \leq M\|x\|$ for all $x \in E$.

{Hint: (a) \Rightarrow (b): Obvious.

(b) \Rightarrow (c): If the set is not bounded, then there exists a sequence (y_n) in E such that $\|y_n\| \leq 1$ and $|f(y_n)| > n$ for all n . Let $x_n = (1/n)y_n$ and contemplate $x_n \rightarrow 0$.

(c) \Rightarrow (d): If $x \neq 0$ then $\|x\|^{-1}x$ has norm 1.

(d) \Rightarrow (a): $|f(x_n) - f(x)| = |f(x_n - x)| \leq M\|x_n - x\|$.

3. Let E be a normed space, f a linear form on E .

(i) If f is continuous, then the number

$$\sup\{|f(x)| : x \in E, \|x\| \leq 1\}$$

is the smallest number $M \geq 0$ satisfying condition (d) in Exercise 2. It is denoted $\|f\|$ and is called the *norm* of f (the terminology is justified in Exercise 4).

(ii) It can be shown that if $a \in E$ and $a \neq 0$, then there exists a continuous linear form f on E such that $\|f\| = 1$ and $f(a) = \|a\|$;⁴ in particular, every nonzero normed space admits nonzero continuous linear forms.

4. Let E be a normed space.

(i) If f and g are continuous linear forms on E and if c is a scalar, then the pointwise sum $f + g$ and scalar multiple cf , defined by the formulas

$$(f + g)(x) = f(x) + g(x), \quad (cf)(x) = cf(x) \quad \text{for all } x \in E,$$

are also continuous, and

$$\|f + g\| \leq \|f\| + \|g\|, \quad \|cf\| = |c| \|f\|.$$

(ii) The set E' of all continuous linear forms on E is a vector space for the pointwise linear operations, and the correspondence $f \mapsto \|f\|$ defines a norm on E' , thus E' is a normed space (over the same field of scalars— \mathbb{R} or \mathbb{C} —as E). In fact, E' is a Banach space (even if E is not complete), called the *dual space* of E .⁵

5. Let (X, \mathcal{S}, μ) be a measure space, let $1 < p < +\infty$, and let $q = p/(p-1)$. For each pair $u \in L^p_{\mathbb{C}}$ and $v \in L^q_{\mathbb{C}}$, define a complex number $\langle u, v \rangle$ as follows. Write $u = \dot{f} = f + \mathcal{N}$ and $v = \dot{g} = g + \mathcal{N}$ (caution: these are cosets in different quotient spaces!) with $f \in \mathcal{L}^p_{\mathbb{C}}$ and $g \in \mathcal{L}^q_{\mathbb{C}}$,

⁴ Cf. the author, *Lectures in functional analysis and operator theory* [Springer-Verlag, New York, 1974], p. 169, 40.10.

⁵ Cf. the author, *op. cit.*, p. 169, 40.9.

and define

$$\langle u, v \rangle = \int f g d\mu;$$

the definition is legitimate because the expression $\int f g d\mu$ depends only on the cosets u and v , not on the particular functions f and g selected to represent them.

(i) The mapping $(u, v) \mapsto \langle u, v \rangle$ is *bilinear*:

$$\begin{aligned}\langle u_1 + u_2, v \rangle &= \langle u_1, v \rangle + \langle u_2, v \rangle \\ \langle cu, v \rangle &= c\langle u, v \rangle \\ \langle u, v_1 + v_2 \rangle &= \langle u, v_1 \rangle + \langle u, v_2 \rangle \\ \langle u, cv \rangle &= c\langle u, v \rangle\end{aligned}$$

for all $u, u_1, u_2 \in L_{\mathbb{C}}^p$, $v, v_1, v_2 \in L_{\mathbb{C}}^q$ and $c \in \mathbb{C}$.

(ii) $|\langle u, v \rangle| \leq \|u\|_p \|v\|_q$.

(iii) For each $v \in L_{\mathbb{C}}^q$, the formula

$$L_v(u) = \langle u, v \rangle \quad (u \in L_{\mathbb{C}}^p)$$

defines a continuous linear form L_v on $L_{\mathbb{C}}^p$, that is, $L_v \in (L_{\mathbb{C}}^p)'$.

(iv) The mapping $L_{\mathbb{C}}^q \rightarrow (L_{\mathbb{C}}^p)'$ defined by $v \mapsto L_v$ is linear:

$$L_{v_1+v_2} = L_{v_1} + L_{v_2}, \quad L_{cv} = cL_v;$$

and isometric: $\|L_v\| = \|v\|_q$.

(v) Theorem 6.7.11 shows that the mapping $v \mapsto L_v$ of (iv) is surjective, assuming μ is finite. In fact, the same is true for an arbitrary measure μ .⁶ The norm-preserving vector space isomorphism $L_{\mathbb{C}}^q \rightarrow (L_{\mathbb{C}}^p)'$ thus defined is usually expressed by writing $(L_{\mathbb{C}}^p)' = L_{\mathbb{C}}^q$. In turn, $(L_{\mathbb{C}}^q)' = L_{\mathbb{C}}^p$, whence $L_{\mathbb{C}}^p = (L_{\mathbb{C}}^p)''$, a property of $L_{\mathbb{C}}^p$ (for $1 < p < +\infty$) known as *reflexivity*.

(vi) If one defines instead $\langle u, v \rangle = \int f \bar{g} d\mu$,⁷ then the correspondence $(u, v) \mapsto \langle u, v \rangle$ becomes *sesquilinear* (linear in u , conjugate-linear in v), and $v \mapsto L_v$ is a conjugate-linear mapping ($L_{cv} = \bar{c}L_v$) of $L_{\mathbb{C}}^q$ onto $(L_{\mathbb{C}}^p)'$. Linearity can be restored in two ways:

(a) consider instead the mapping $v \mapsto L_{\bar{v}}$ where, for $v = g + \mathcal{N} \in L_{\mathbb{C}}^q$, one defines $\bar{v} = \bar{g} + \mathcal{N}$, \bar{g} being the complex-conjugate function $\bar{g}(x) = \overline{g(x)}$ ($x \in X$); or

(b) stick to $v \mapsto L_v$ but replace the natural (pointwise) scalar multiple $(c, L) \mapsto cL$ on $(L_{\mathbb{C}}^p)'$ by the scalar multiple $(c, L) \mapsto \bar{c}L$.

If $p = q = 2$ then $\langle u, v \rangle$ is defined for all $u, v \in L_{\mathbb{C}}^2$ (called the *inner product*, or *scalar product*, of u and v), in particular $\langle u, u \rangle = (\|u\|_2)^2$;

⁶ Cf. H.L. Royden, *op. cit.*, p. 286, Theorem 30.

⁷ E. Hewitt and K. Stromberg, *Real and abstract analysis* [Springer-Verlag, New York, 1965], p. 223, (15.1).

$\mathcal{L}_{\mathbb{C}}^2$ is an example of a (complex) *Hilbert space*, that is, a Banach space whose norm satisfies the ‘parallelogram law’

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

for all u and v .⁸

6. (i) If (X, \mathcal{S}, μ) is a *finite* measure space and $1 \leq p < r$, then $\mathcal{L}_{\mathbb{C}}^r(\mu) \subset \mathcal{L}_{\mathbb{C}}^p(\mu)$.

(ii) If μ is not finite, the inclusion in (i) is in general false.

(iii) If $f : [0, 1] \rightarrow \mathbb{R}$ is the function defined by $f(0) = 0$ and $f(x) = x^{-1/2}$ for $0 < x \leq 1$, then f is Lebesgue-integrable but its square is not.

(iv) The measure space in the Hint for §6.6, Exercise 2 is not finite, but all of the spaces \mathcal{L}^p ($1 \leq p < +\infty$) coincide.

{Hint: (i) If $f \in \mathcal{L}_{\mathbb{C}}^r$ and

$$E = \{x : |f(x)| \leq 1\}, \quad F = \{x : |f(x)| > 1\},$$

then $\varphi_E f \in \mathcal{L}_{\mathbb{C}}^{\infty} \subset \mathcal{L}_{\mathbb{C}}^p$ and $|\varphi_F f|^p \leq |\varphi_F f|^r \in \mathcal{L}^1$; contemplate $f = \varphi_E f + \varphi_F f$. (ii) Cf. §3.1, Exercise 4.}

6.8. $\mathcal{C}(X)$

In the preceding sections, the Banach spaces of functions on a set X have come from imposing a measure-theoretic structure on X and requiring the functions to be measurable. In the present section, the structure imposed on X is topological, the functions are required to be continuous, and the key results are obtained when X is assumed to be quasicompact; much of the ground has already been prepared in our discussion of uniform convergence (§6.2, especially 6.2.23). We begin with a discussion of algebraic operations in the space $\mathcal{C}_{\mathbb{R}}(X)$ of continuous real-valued functions on X (6.2.20), then enlarge the discussion to accommodate complex-valued functions. The major applications come in the next section: Weierstrass’s polynomial approximation theorem for continuous functions on a closed interval, and its spectacular generalization by M.H. Stone for continuous functions on a compact space.

6.8.1. Lemma. *For a real-valued function $f : X \rightarrow \mathbb{R}$ on a topological space X , the following conditions are equivalent:*

- (a) f is continuous;
- (b) for every real number c , the sets

$$\{x \in X : f(x) < c\}, \quad \{x \in X : f(x) > c\}$$

are open sets in X ;

⁸Hewitt and Stromberg, *op. cit.*, p. 235, (16.8); or the author, *op. cit.*, p. 164, (39.10) and p. 174, (41.1).

(c) for every real number c , the sets

$$\{x \in X : f(x) \geq c\}, \quad \{x \in X : f(x) \leq c\}$$

are closed sets in X .

Proof. (a) \Rightarrow (b): The sets

$$\{x : f(x) < c\} = f^{-1}((-\infty, c)), \quad \{x : f(x) > c\} = f^{-1}((c, +\infty))$$

are the inverse images of open sets in \mathbb{R} , hence are open in X (3.4.5).

(b) \Rightarrow (a): For every open interval (a, b) in \mathbb{R} , its inverse image

$$\begin{aligned} f^{-1}((a, b)) &= \{x : a < f(x) < b\} \\ &= \{x : f(x) > a\} \cap \{x : f(x) < b\} \end{aligned}$$

is open in X (being the intersection of two open sets). If U is any open set in \mathbb{R} , then U is the union of a family of open intervals, say $U = \bigcup_{i \in I} (a_i, b_i)$; then

$$f^{-1}(U) = \bigcup_{i \in I} f^{-1}((a_i, b_i))$$

is the union of a family of open sets in X , hence is open. Thus the inverse image of every open set in \mathbb{R} is open in X , therefore f is continuous (3.4.5).

(b) \Leftrightarrow (c): The sets described in (c) are the complements of the sets described in (b), thus the equivalence is immediate from the definition of closed set (3.3.1). \diamond

6.8.2. Proposition. *If f and g are continuous real-valued functions on a topological space X , and if $a \in \mathbb{R}$ and $\alpha > 0$, then the following real-valued functions on X are also continuous:*

$$f + g, \quad af, \quad fg, \quad |f|^\alpha, \quad f \cup g, \quad f \cap g, \quad f^+, \quad f^-;$$

if, moreover, $f(x) \neq 0$ ($\forall x \in X$), then $1/f$ is continuous.

In particular, $C_{\mathbb{R}}(X)$ is an algebra over \mathbb{R} (for the pointwise operations $f + g, af, fg$), containing the constant functions.

Proof. The function $x \mapsto (af)(x) = af(x)$ is the composite of the continuous functions $x \mapsto f(x)$ and $r \mapsto ar$ ($r \in \mathbb{R}$), hence is continuous.

For every real number c , the set

$$\begin{aligned} \{x : (f + g)(x) < c\} &= \{x : f(x) < -g(x) + c\} \\ &= \bigcup_{r \in \mathbb{R}} \{x : f(x) < r < -g(x) + c\} \\ &= \bigcup_{r \in \mathbb{R}} \{x : f(x) < r\} \cap \{x : g(x) < c - r\} \end{aligned}$$

is open in X by the lemma; in view of the preceding paragraph, so is the set

$$\{x : (f + g)(x) > c\} = \{x : ((-f) + (-g))(x) < -c\},$$

thus $f + g$ is continuous by the lemma.

The function $x \mapsto (|f|^\alpha)(x) = |f(x)|^\alpha$ is the composite of the continuous functions $x \mapsto f(x)$ and $r \mapsto |r|^\alpha$ ($r \in \mathbb{R}$), hence is continuous. In particular, the functions $|f|$ and $f^2 = |f|^2$ are continuous, therefore so are the functions

$$\begin{aligned} fg &= \frac{1}{4}[(f + g)^2 - (f - g)^2] \\ f \cup g &= \frac{1}{2}(f + g + |f - g|) \\ f \cap g &= \frac{1}{2}(f + g - |f - g|) \\ f^+ &= f \cup 0, \quad f^- = -(f \cap 0). \end{aligned}$$

Finally, if $f(x)$ is never 0 then $1/f$ is the composite of the continuous mappings $x \mapsto f(x)$ and $r \mapsto 1/r$ ($r \in \mathbb{R} - \{0\}$). \diamond

6.8.3. We now advance to complex-valued functions on a topological space X . The objective is to study $\mathcal{C}_{\mathbb{C}}(X)$ by applying what we know about $\mathcal{C}_{\mathbb{R}}(X)$ to the real and imaginary parts of functions $f \in \mathcal{C}_{\mathbb{C}}(X)$.

The first step is to establish that $\mathcal{C}_{\mathbb{R}}(X)$ can be regarded as a subset of $\mathcal{C}_{\mathbb{C}}(X)$, namely that

$$(*) \quad \mathcal{C}_{\mathbb{R}}(X) = \{f \in \mathcal{C}_{\mathbb{C}}(X) : f \text{ is real-valued}\}.$$

There are three technical obstacles to asserting this “obvious” formula (easily overcome, as we shall see): (i) the functions belonging to the left side are functions $f : X \rightarrow \mathbb{R}$, whereas those belonging to the right side are functions $f : X \rightarrow \mathbb{C}$ for which $f(X) \subset \mathbb{R}$, and we have pledged that when two functions are regarded as “equal”, they should in particular have the same final set (1.3.1); (ii) if $f \in \mathcal{C}_{\mathbb{R}}(X)$, we know that the inverse image under f of an open set in \mathbb{R} is open in X , but to show that the function $X \rightarrow \mathbb{C}$ having the same graph as f is continuous, we must contemplate inverse images of open sets in \mathbb{C} , the rub being that the only set that is open in both \mathbb{R} and \mathbb{C} is the empty set; (iii) if $f \in \mathcal{C}_{\mathbb{C}}(X)$ and $f(X) \subset \mathbb{R}$, we know that the inverse image of an open set in \mathbb{C} is open in X , but to regard f as a continuous function $X \rightarrow \mathbb{R}$ we must contemplate inverse images of open sets in \mathbb{R} .

The obstacles (i)–(iii) feel like three variations on the same obstacle; in fact, by fixing (i) we can fix all three. If $g : X \rightarrow \mathbb{R}$ and $\iota : \mathbb{R} \rightarrow \mathbb{C}$ is the insertion mapping $\iota(r) = r$ ($r \in \mathbb{R}$), then the composite mapping $f = \iota \circ g$ is a mapping $f : X \rightarrow \mathbb{C}$ having the same graph as g , such that $f(X) = g(X) \subset \mathbb{R}$. We thus have a mapping $g \mapsto \iota \circ g$ (evidently

injective) of $\mathcal{F}(X, \mathbb{R})$ into $\mathcal{F}(X, \mathbb{C})$.¹ If, conversely, $f : X \rightarrow \mathbb{C}$ and $f(X) \subset \mathbb{R}$, and if $\text{Re} : \mathbb{C} \rightarrow \mathbb{R}$ is the projection mapping $\text{Re}(a + bi) = a$ ($a, b \in \mathbb{R}$), then the composite $g = \text{Re} \circ f$ is a function $g : X \rightarrow \mathbb{R}$ such that $\iota \circ g = f$.

Conclusion:

$$\{\iota \circ g : g \in \mathcal{F}(X, \mathbb{R})\} = \{f \in \mathcal{F}(X, \mathbb{C}) : f(X) \subset \mathbb{R}\},$$

and the mapping $g \mapsto \iota \circ g$ is a bijection

$$\mathcal{F}(X, \mathbb{R}) \rightarrow \{f \in \mathcal{F}(X, \mathbb{C}) : f(X) \subset \mathbb{R}\}$$

with inverse mapping $f \mapsto \text{Re} \circ f$. The wish expressed in (*) should therefore be revised to the wish that

$$(**) \quad \{\iota \circ g : g \in \mathcal{C}_{\mathbb{R}}(X)\} = \{f \in \mathcal{C}_{\mathbb{C}}(X) : f(X) \subset \mathbb{R}\};$$

once (**) is verified, we can safely identify $\mathcal{C}_{\mathbb{R}}(X)$ with the set of real-valued functions in $\mathcal{C}_{\mathbb{C}}(X)$. We prepare the way with three easy lemmas.

6.8.4. Lemma. View \mathbb{R} as a subset of \mathbb{C} in the usual way ($\mathbb{R} = \{r + 0i : r \in \mathbb{R}\}$), each of \mathbb{R} and \mathbb{C} being equipped with the (usual) absolute-value metric (3.2.2). Then

$$\mathcal{O}_{\mathbb{R}} = \{U \cap \mathbb{R} : U \in \mathcal{O}_{\mathbb{C}}\},$$

concisely $\mathcal{O}_{\mathbb{R}} = \mathcal{O}_{\mathbb{C}} \cap \mathbb{R}$ (where \mathcal{O} denotes the class of open sets for the topology in question).

Proof. The proof is a double inclusion argument.

\subset : Let A be an open set in \mathbb{R} ; we seek an open set U in \mathbb{C} such that $A = U \cap \mathbb{R}$. The set $U = A + \mathbb{R}i = \{z \in \mathbb{C} : \text{Re } z \in A\}$ satisfies $U \cap \mathbb{R} = A$. Moreover, U is open in \mathbb{C} ; for, if $c \in U$, say $c = a + ri$ with $a \in A$ and $r \in \mathbb{R}$, and if $\epsilon > 0$ is chosen so that $(a - \epsilon, a + \epsilon) \subset A$, then the inclusion $\{z \in \mathbb{C} : |z - c| < \epsilon\} \subset U$ follows from the fact that $|\text{Re}(z) - a| = |\text{Re}(z) - \text{Re}(c)| = |\text{Re}(z - c)| \leq |z - c|$.

\supset : Let U be an open set in \mathbb{C} and let $A = U \cap \mathbb{R}$; we are to show that A is open in \mathbb{R} . Thus, given any $a \in A$, it suffices to find an $\epsilon > 0$ such that $(a - \epsilon, a + \epsilon) \subset A$. Since a also belongs to U , hence is an interior point of U in \mathbb{C} , there exists an $\epsilon > 0$ such that the disc $D = \{c \in \mathbb{C} : |c - a| < \epsilon\}$ is contained in U , therefore $(a - \epsilon, a + \epsilon) = D \cap \mathbb{R} \subset U \cap \mathbb{R} = A$. \diamond

6.8.5. Lemma. If $\iota : \mathbb{R} \rightarrow \mathbb{C}$ is the insertion mapping $\iota(r) = r$ ($r \in \mathbb{R}$), then

$$\mathcal{O}_{\mathbb{R}} = \iota^{-1}(\mathcal{O}_{\mathbb{C}}) = \{\iota^{-1}(U) : U \in \mathcal{O}_{\mathbb{C}}\}.$$

In particular, ι is continuous.

¹ In general, if X and Y are nonempty sets, then $\mathcal{F}(X, Y)$ denotes the set of all functions $f : X \rightarrow Y$ (1.13.10), with equality in $\mathcal{F}(X, Y)$ defined as in 1.3.1.

Proof. Since $\iota^{-1}(U) = U \cap \mathbb{R}$ for every subset U of \mathbb{R} , the asserted equality is just a restatement of the preceding lemma. In particular, $\iota^{-1}(\mathcal{O}_{\mathbb{C}}) \subset \mathcal{O}_{\mathbb{R}}$, so ι is continuous. \diamond

6.8.6. Lemma. *If $g : X \rightarrow \mathbb{R}$ is a real-valued function on a topological space X , and $\iota : \mathbb{R} \rightarrow \mathbb{C}$ is the insertion mapping, then*

$$g : X \rightarrow \mathbb{R} \text{ is continuous} \Leftrightarrow \iota \circ g : X \rightarrow \mathbb{C} \text{ is continuous}$$

Proof. \Rightarrow : Immediate from the fact that ι is continuous (6.8.5).

\Leftarrow : If A is an open set in \mathbb{R} , we assert that $g^{-1}(A)$ is open in X . Choose an open set U in \mathbb{C} such that $A = U \cap \mathbb{R}$ (6.8.4). By assumption, $(\iota \circ g)^{-1}(U)$ is open in X ; but $(\iota \circ g)^{-1}(U) = g^{-1}(\iota^{-1}(U)) = g^{-1}(U \cap \mathbb{R}) = g^{-1}(A)$, whence the assertion. \diamond

With the foregoing lemmas in hand, the proof of $(**)$ is effortless:

6.8.7. Lemma. *If X is a topological space and $\iota : \mathbb{R} \rightarrow \mathbb{C}$ is the insertion mapping, then*

$$\{\iota \circ g : g \in \mathcal{C}_{\mathbb{R}}(X)\} = \{f \in \mathcal{C}_{\mathbb{C}}(X) : f(X) \subset \mathbb{R}\}.$$

Proof. The proof is a double inclusion argument.

\subset : Immediate from the preceding lemma.

\supset : If $f \in \mathcal{C}_{\mathbb{C}}(X)$ and $f(X) \subset \mathbb{R}$, write g for the function $X \rightarrow \mathbb{R}$ having the graph of f ; since $\iota \circ g = f$ is known to be continuous, so is g by the preceding lemma. \diamond

The preceding lemmas involve the insertion mapping $\iota : \mathbb{R} \rightarrow \mathbb{C}$; in the reverse direction $\mathbb{C} \rightarrow \mathbb{R}$, we have the two coordinate projections:

6.8.8. Definition. We write $\text{Re} : \mathbb{C} \rightarrow \mathbb{R}$ and $\text{Im} : \mathbb{C} \rightarrow \mathbb{R}$ for the functions (clearly \mathbb{R} -linear mappings)

$$\text{Re}(a + bi) = a, \quad \text{Im}(a + bi) = b$$

for all $a, b \in \mathbb{R}$. If $f : X \rightarrow \mathbb{C}$ is a complex-valued function on a set X , we write $\text{Re } f$ and $\text{Im } f$ for the composite functions $\text{Re} \circ f : X \rightarrow \mathbb{R}$ and $\text{Im} \circ f : X \rightarrow \mathbb{R}$, called the **real part** and the **imaginary part** of f .

6.8.9. Lemma. *If $f : X \rightarrow \mathbb{C}$ is a complex-valued function on a topological space X , then*

$$f : X \rightarrow \mathbb{C} \text{ is continuous} \Leftrightarrow \text{Re } f, \text{Im } f : X \rightarrow \mathbb{R} \text{ are continuous.}$$

Proof. If A is an open set in \mathbb{R} , then the set

$$\text{Re}^{-1}(A) = \{a + bi : a \in A, b \in \mathbb{R}\} = A + \mathbb{R}i$$

is open in \mathbb{C} by the proof of 6.8.4, thus Re is continuous. The proof for Im is similar.

Consider \mathbb{R}^2 with its usual topology, derived from its Euclidean metric

$$d_2((a, b), (a', b')) = [(a - a')^2 + (b - b')^2]^{1/2}$$

(cf. 3.1.15); this topology is also generated by the sup-metric

$$d_\infty((a, b), (a', b')) = \max\{|a - a'|, |b - b'|\}$$

(cf. 3.3.7, with \mathbb{C}^r replaced by \mathbb{R}^r). The bijective mapping

$$\theta : \mathbb{R}^2 \rightarrow \mathbb{C}$$

defined by $\theta(a, b) = a + bi$ is isometric (distance-preserving) for the metric d_2 on \mathbb{R}^2 and the absolute-value metric on \mathbb{C} (3.1.9), hence is a homeomorphism for the corresponding topologies.

Suppose $f : X \rightarrow \mathbb{C}$ and write $g = \operatorname{Re} f$, $h = \operatorname{Im} f$. If f is continuous, then so are the composite functions $\operatorname{Re} \circ f = g$ and $\operatorname{Im} \circ f = h$.

Conversely, assuming that g and h are continuous, we are to show that f is continuous. Define a mapping $F : X \rightarrow \mathbb{R}^2$ by the formula

$$F(x) = (g(x), h(x)) \quad (x \in X).$$

Consider an 'open ball' U in (\mathbb{R}^2, d_∞) , say

$$\begin{aligned} U &= U_r(a_0, b_0) = \{(a, b) : d_\infty((a, b), (a_0, b_0)) < r\} \\ &= \{(a, b) : \max(|a - a_0|, |b - b_0|) < r\} \\ &= \{(a, b) : |a - a_0| < r\} \cap \{(a, b) : |b - b_0| < r\} \\ &= [(a_0 - r, a_0 + r) \times \mathbb{R}] \cap [\mathbb{R} \times (b_0 - r, b_0 + r)] \end{aligned}$$

(note, incidentally, that U is a square for the metric d_2); then the set

$$\begin{aligned} F^{-1}(U) &= \{x : ((g(x), h(x)) \in U\} \\ &= \{x : |g(x) - a_0| < r\} \cap \{x : |h(x) - b_0| < r\} \\ &= g^{-1}((a_0 - r, a_0 + r)) \cap h^{-1}((b_0 - r, b_0 + r)) \end{aligned}$$

is open in X (by the assumed continuity of g and h). Since every open set in \mathbb{R}^2 is the union of such open balls U , we conclude that F is continuous, therefore so is the composite function $\theta \circ F = f$. \diamond

We are ready at last to discuss efficiently the algebraic operations on continuous complex-valued functions:

6.8.10. Proposition. *If f and g are continuous complex-valued functions on a topological space X , and if $c \in \mathbb{C}$ and $\alpha > 0$, then the following functions on X are also continuous:*

$$f + g, \quad cf, \quad fg, \quad \operatorname{Re} f, \quad \operatorname{Im} f, \quad \bar{f}, \quad |f|^\alpha;$$

if, moreover, $f(x) \neq 0$ ($\forall x \in X$), then $1/f$ is continuous. In particular, $\mathcal{C}_{\mathbb{C}}(X)$ is an algebra (for the pointwise operations $f+g, cf, fg$) containing the constant functions.

Proof. Recall that the functions $\operatorname{Re} f$ and $\operatorname{Im} f$ have final set \mathbb{R} (6.8.8); for the purposes of this proof, we assume that $|f|^\alpha$ also has final set \mathbb{R} . After the proof is done, we shall relax all this fussiness over final sets (6.8.11).

We know from the preceding lemma that the real and imaginary parts of f and g are continuous (i.e., belong to $\mathcal{C}_{\mathbb{R}}(X)$), therefore so are the functions $\operatorname{Re}(f+g) = \operatorname{Re} f + \operatorname{Re} g$ and $\operatorname{Im}(f+g) = \operatorname{Im} f + \operatorname{Im} g$ (6.8.2), consequently $f+g \in \mathcal{C}_{\mathbb{C}}(X)$ (again by the preceding lemma). Similarly, $\operatorname{Re} \bar{f} = \operatorname{Re} f$ and $\operatorname{Im} \bar{f} = -\operatorname{Im} f$ are continuous, therefore so is \bar{f} . The proofs for cf and $\bar{f}g$ are similar.

In particular, $f\bar{f}$ is continuous, therefore so is $|f|^\alpha = (f\bar{f})^{\alpha/2}$ by Proposition 6.8.2 (note that, in view of Lemma 6.8.6, it does not matter whether we regard \mathbb{R} or \mathbb{C} as the final set for $f\bar{f}$ and $|f|^\alpha$). Finally, if f is never 0 then $|f|^{-2} = (|f|^2)^{-1}$ is continuous (6.8.2), therefore so is $1/f = |f|^{-2}\bar{f}$. \diamond

The foregoing results (particularly 6.8.6) amply justify the following simplification:

6.8.11. *Scholium.* With notations as in the preceding proposition, we regard $\mathcal{C}_{\mathbb{R}}(X)$ as an \mathbb{R} -subalgebra of $\mathcal{C}_{\mathbb{C}}(X)$, by identifying $g \in \mathcal{C}_{\mathbb{R}}(X)$ with $\iota \circ g \in \mathcal{C}_{\mathbb{C}}(X)$ (see 6.8.6). In particular, for $f \in \mathcal{C}_{\mathbb{C}}(X)$, the functions $\operatorname{Re} f$, $\operatorname{Im} f$ and $|f|^\alpha$ are henceforth regarded as continuous functions $X \rightarrow \mathbb{C}$ (whose values happen to be in \mathbb{R}), and the formulas $\operatorname{Re} f = \frac{1}{2}f + \frac{1}{2}\bar{f}$, $f = \operatorname{Re} f + i\operatorname{Im} f$, etc., may be regarded as linear combinations formed in the complex vector space $\mathcal{C}_{\mathbb{C}}(X)$.

When X is quasicompact, $\mathcal{C}_{\mathbb{C}}(X)$ has in addition a natural Banach space structure:

6.8.12. Theorem. Let X be a quasicompact topological space (6.1.3). Then:

- (i) Every continuous function $f: X \rightarrow \mathbb{C}$ is bounded.
- (ii) Writing $\|f\|_\infty = \sup_{x \in X} |f(x)|$ for $f \in \mathcal{C}_{\mathbb{C}}(X)$, the correspondence $f \mapsto \|f\|_\infty$ defines a norm (6.4.9) on the complex vector space $\mathcal{C}_{\mathbb{C}}(X)$.
- (iii) $\mathcal{C}_{\mathbb{C}}(X)$ is a complete metric space for the sup-metric $(f, g) \mapsto \|f - g\|_\infty$, thus is a Banach space (6.4.17).
- (iv) $\|fg\|_\infty \leq \|f\|_\infty \|g\|_\infty$ for all $f, g \in \mathcal{C}_{\mathbb{C}}(X)$.
- (v) $\|f\bar{f}\|_\infty = (\|f\|_\infty)^2$ for all $f \in \mathcal{C}_{\mathbb{C}}(X)$.

Proof. (i) See Proposition 6.2.11.

(ii) The property $\|f+g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$ (called ‘subadditivity’) is verified in 3.1.10, and it is easy to verify that $\|cf\|_\infty = |c| \|f\|_\infty$ (‘absolute homogeneity’) and that $\|f\|_\infty > 0$ unless f is the function identically zero (‘strict positivity’), thus $f \mapsto \|f\|_\infty$ is a norm on $\mathcal{C}_{\mathbb{C}}(X)$ (justifying the term “sup-norm” already employed in 3.1.10).

(iii) Completeness for the sup-norm metric is noted in 6.2.23, thus $C_{\mathbb{C}}(X)$ is a (complex) Banach space.

(iv) For all $x \in X$,

$$|(fg)(x)| = |f(x)g(x)| = |f(x)| \cdot |g(x)| \leq \|f\|_{\infty} \|g\|_{\infty},$$

therefore $\|fg\|_{\infty} \leq \|f\|_{\infty} \|g\|_{\infty}$.

(v) The asserted equality is a restatement of the equality

$$\sup_{x \in X} |f(x)|^2 = \left(\sup_{x \in X} |f(x)| \right)^2,$$

valid because $r \mapsto r^2$ is an order isomorphism $[0, +\infty) \rightarrow [0, +\infty)$. \diamond

6.8.13. *Definition.* A Banach space (real or complex) with an associative bilinear multiplication $(a, b) \mapsto ab$ satisfying $\|ab\| \leq \|a\| \|b\|$ is called a **Banach algebra** (real or complex).²

6.8.14. Theorem 6.8.12 shows that $C_{\mathbb{C}}(X)$ is a complex Banach algebra. The analogous theorem for real-valued functions holds, showing that $C_{\mathbb{R}}(X)$ is a real Banach algebra; property (v) then takes the form $\|f^2\|_{\infty} = (\|f\|_{\infty})^2$. The identity (v) (in both the real and complex cases) is characteristic of the so-called *C*-algebras*.³

Exercises

1. The message of Lemma 6.8.4 is that the usual topology on \mathbb{R} is the relative topology induced on it by the usual topology on \mathbb{C} (cf. §3.3, Exercise 7). Extend this result to the relative topology on a subset of a metric space.

2. Let X and Y be nonempty topological spaces, $Z = X \times Y$ the product set. Call a subset W of Z *open* if, for each point $z \in W$, there exist open sets U and V in X and Y , respectively, such that $z \in U \times V \subset W$, and let \mathcal{O}_Z be the set of all such subsets W . Prove:

(i) \mathcal{O}_Z is the family of open sets for a topology on Z , that is, (Z, \mathcal{O}_Z) is a topological space in the sense of Definition 3.3.1; \mathcal{O}_Z is called the *product topology* on Z , and (Z, \mathcal{O}_Z) is called the *product topological space* (of the topological spaces X and Y , in that order).

Assume for the rest of the exercise that Z is equipped with the product topology.

(ii) Let $\text{pr}_X : Z \rightarrow X$ and $\text{pr}_Y : Z \rightarrow Y$ be the projection mappings,

² C.E. Rickart, *General theory of Banach algebras* [Van Nostrand, Princeton, NJ, 1960; reprinted by R.E. Krieger, Huntington, NY, 1974].

³ Cf. J. Dixmier, *C*-algebras* [North-Holland, Amsterdam, 1977], R.V. Kadison and J.R. Ringrose, *Fundamentals of the theory of operator algebras*, Vols. I–IV [Academic Press, New York, 1983–1992].

defined by

$$\text{pr}_X(x, y) = x, \quad \text{pr}_Y(x, y) = y.$$

For subsets $A \subset X$ and $B \subset Y$, $\text{pr}_X^{-1}(A) = A \times Y$ and $\text{pr}_Y^{-1}(B) = X \times B$; infer that pr_X and pr_Y are continuous mappings.

(iii) If W is an open set in Z , then $\text{pr}_X(W)$ and $\text{pr}_Y(W)$ are open sets in X and Y , respectively.

(iv) If f is a function defined on a topological space T and taking values in the product topological space $Z = X \times Y$, then $f : T \rightarrow Z$ is continuous if and only if both $\text{pr}_X \circ f : T \rightarrow X$ and $\text{pr}_Y \circ f : T \rightarrow Y$ are continuous, concisely,

$$f : T \rightarrow X \times Y \text{ continuous} \Leftrightarrow \text{pr}_X \circ f \text{ and } \text{pr}_Y \circ f \text{ continuous.}$$

3. In contrast with Lemma 6.8.9, why was it not necessary in §6.4 to struggle to show that a complex-valued function on a measurable space is measurable if and only if its real and imaginary parts are measurable?

4. The set $\mathcal{BV}[a, b]$ of all functions $f : [a, b] \rightarrow \mathbb{R}$ of bounded variation, equipped with the pointwise linear operations and product, is a (real) Banach algebra for the norm $\|f\|' = \|f\|_\infty + V_a^b f$ (§5.1, Exercise 4).⁴

5. Let T be a nonempty set, and let $\mathcal{B}_\mathbb{C}(T)$ be the complex vector space of all bounded, complex-valued functions on T , equipped with the pointwise linear operations and the sup-norm (3.1.10). With products in $\mathcal{B}_\mathbb{C}(T)$ defined pointwise, $\mathcal{B}_\mathbb{C}(T)$ is a complex Banach algebra. Similarly, $\mathcal{B}_\mathbb{R}(T)$ is a real Banach algebra for the pointwise operations and the sup-norm.

{Hint: Example 6.2.17 and the computation in (iv) of Theorem 6.8.12.}

6.9. Stone-Weierstrass Approximation Theorem

If $\sum_{k=0}^{\infty} c_k t^k$ is a power series with real coefficients and radius of convergence $R > 0$ (§1.16, Exercise 3) and if $[a, b]$ is a nondegenerate closed subinterval of $(-R, R)$, then the series converges uniformly and absolutely on $[a, b]$ (Example 6.2.8); the formula

$$f(t) = \sum_{k=0}^{\infty} c_k t^k \quad (a \leq t \leq b)$$

defines a function $f : [a, b] \rightarrow \mathbb{R}$ that is continuous on $[a, b]$ and differentiable on (a, b) (Example 6.2.28). In particular, f is the uniform limit on $[a, b]$ of a sequence of polynomial functions (the sequence of functions defined by the partial sums).

⁴ Cf. C.E. Rickart, *op. cit.*, p. 302, A.2.5.

Not every continuous function $g : [a, b] \rightarrow \mathbb{R}$ has such a power series representation; for example, the continuous function $g : [-1, 1] \rightarrow \mathbb{R}$ defined by $g(t) = |t|$ fails to be differentiable at the origin. Nevertheless, Weierstrass¹ proved that every continuous function $g : [a, b] \rightarrow \mathbb{R}$ is the uniform limit of a sequence of polynomial functions (we just can't expect the differences of successive terms of the sequence to be monomials of increasing degree). Amazingly, the crux of the matter is to prove that the function $g(t) = |t|$ on $[-1, 1]$ is such a uniform limit.

Stated in topological terms, Weierstrass's theorem says that in the algebra $C_{\mathbb{R}}[a, b]$, equipped with the metric defined by the sup-norm (6.2.23), the subalgebra consisting of the polynomial functions is a dense subset. In a *tour-de-force* of analysis, M.H. Stone² isolated the key elements of the proof of Weierstrass' theorem and recast them in a vastly more general theorem about the approximation of continuous functions on a compact space X , the algebra of polynomial functions being replaced by a suitable subalgebra of $C_{\mathbb{R}}(X)$.³ The present section is devoted to an exposition of Stone's theorem.

Notations fixed for the rest of the section: X is a compact topological space (6.1.6); as in the preceding section, $C_{\mathbb{R}}(X)$ and $C_{\mathbb{C}}(X)$ are the algebras of real-valued and complex-valued continuous functions on X , equipped with the pointwise operations and the sup-metric.

The core result is a theorem about linear subspaces of $C_{\mathbb{R}}(X)$:

6.9.1. Theorem. *If \mathcal{L} is a linear subspace of $C_{\mathbb{R}}(X)$ such that*

1° \mathcal{L} *separates the points of* X ,

2° \mathcal{L} *annihilates no point of* X , and

3° $f \in \mathcal{L} \Rightarrow f \cap 1 \in \mathcal{L}$,

then \mathcal{L} is dense in $C_{\mathbb{R}}(X)$ for the sup-metric, that is, every $f \in C_{\mathbb{R}}(X)$ is the uniform limit of a sequence of functions in \mathcal{L} .

Before embarking on the proof, which is divided into a series of five lemmas, some comments on the conditions 1°–3° are in order.

6.9.2. Remarks. 1. The meaning of 1°: If $x, y \in X$ with $x \neq y$, then there exists a function $f \in \mathcal{L}$ such that $f(x) \neq f(y)$.

2. The meaning of 2°: For each $x \in X$ there exists a function $f \in \mathcal{L}$ such that $f(x) \neq 0$. (The condition is trivially satisfied if \mathcal{L} contains the constant function 1.)

3. The meaning of 3°: If $f \in \mathcal{L}$ then \mathcal{L} also contains the function obtained by truncating the graph of f from above at 1, that is, the function $(f \cap 1)(x) = \min\{f(x), 1\}$.

¹ Karl Weierstrass (1815–1897).

² Marshall Harvey Stone (1903–1989).

³ M.H. Stone, "The generalized Weierstrass approximation theorem" [*Mathematics Magazine* 21 (1948), 167–184, 237–254].

4. None of the conditions 1°–3° can be omitted in Theorem 6.9.1 (Exercise 1).

6.9.3. Lemma. *With \mathcal{L} as in 6.9.1, the uniform closure of \mathcal{L} is also a linear subspace of $\mathcal{C}_{\mathbb{R}}(X)$ satisfying 1°–3°.*

Proof. Write $\overline{\mathcal{L}}$ for the closure of \mathcal{L} in $\mathcal{C}_{\mathbb{R}} = \mathcal{C}_{\mathbb{R}}(X)$ for the sup-metric. It is obvious that $\overline{\mathcal{L}}$ satisfies 1° and 2°; our problem is to show that $\overline{\mathcal{L}}$ contains sums and scalar multiples and that it satisfies 3°.

Let $f, g \in \overline{\mathcal{L}}$ and choose sequences $(f_n), (g_n)$ in \mathcal{L} such that $f_n \rightarrow f, g_n \rightarrow g$ uniformly. Then $f_n + g_n \rightarrow f + g$ uniformly and, for every $c \in \mathbb{R}$, $cf_n \rightarrow cf$ uniformly; since $f_n + g_n$ and cf_n belong to \mathcal{L} , their uniform limits $f + g$ and cf belong to $\overline{\mathcal{L}}$. Thus $\overline{\mathcal{L}}$ is a linear subspace of $\mathcal{C}_{\mathbb{R}}$. Moreover,

$$f_n \cap 1 = \frac{1}{2}\{f_n + 1 - |f_n - 1|\}, \quad f \cap 1 = \frac{1}{2}\{f + 1 - |f - 1|\};$$

since $f_n \cap 1 \in \mathcal{L}$ and $f_n \cap 1 \rightarrow f \cap 1$ uniformly, we conclude that $f \cap 1 \in \overline{\mathcal{L}}$. \diamond

In view of the preceding lemma, the assertion of Theorem 6.9.1 is that if \mathcal{L} is a *closed* linear subspace of $\mathcal{C}_{\mathbb{R}}$ (for the sup-metric topology) satisfying 1°–3°, then $\mathcal{L} = \mathcal{C}_{\mathbb{R}}$; the next lemma is a small but crucial part of the assertion:

6.9.4. Lemma. *If \mathcal{L} is a closed linear subspace of $\mathcal{C}_{\mathbb{R}}(X)$ satisfying 1°–3°, then*

$$f \in \mathcal{L} \Rightarrow |f| \in \mathcal{L}.$$

Proof. Let $f \in \mathcal{L}$. For every positive integer n , $f \cap \frac{1}{n} = \frac{1}{n}((nf) \cap 1) \in \mathcal{L}$ and

$$f \cap \frac{1}{n} = \frac{1}{2}\{f + \frac{1}{n} - |f - \frac{1}{n}|\} \rightarrow \frac{1}{2}(f - |f|) = f \cap 0$$

uniformly, therefore $f \cap 0 \in \overline{\mathcal{L}} = \mathcal{L}$. Then also $f \cup 0 = -((-f) \cap 0) \in \mathcal{L}$, so $|f| = (f \cup 0) - (f \cap 0) \in \mathcal{L}$. \diamond

6.9.5. Lemma. *If \mathcal{L} is a closed linear subspace of $\mathcal{C}_{\mathbb{R}}(X)$ satisfying 1°–3°, then*

$$f, g \in \mathcal{L} \Rightarrow f \cup g, f \cap g \in \mathcal{L}.$$

Proof. This is immediate from the preceding lemma and the formulas

$$f \cup g = \frac{1}{2}\{f + g + |f - g|\}, \quad f \cap g = \frac{1}{2}\{f + g - |f - g|\}. \quad \diamond$$

6.9.6. Lemma. *If \mathcal{L} is a closed linear subspace of $\mathcal{C}_{\mathbb{R}}(X)$ satisfying 1°–3°, then \mathcal{L} contains the constant functions.*

Proof. We need only show that $1 \in \mathcal{L}$. For each $x \in X$, choose $f_x \in \mathcal{L}$ with $f_x(x) \neq 0$ (possible by 2°). Replacing f_x by $|f_x|$, we can suppose

(Lemma 6.9.4) that $f_x(x) > 0$, $f_x \geq 0$ on X . Multiplying by a scalar, we can further suppose that

$$f_x(x) > 1, \quad f_x \geq 0 \quad \text{on } X.$$

Let

$$U_x = \{y \in X : f_x(y) > 1\} = f_x^{-1}((1, +\infty));$$

then $x \in U_x$ and, by the continuity of f_x , U_x is an open set in X . Thus $(U_x)_{x \in X}$ is an open covering of X . By compactness, there is a finite subcovering

$$X = U_{x_1} \cup \dots \cup U_{x_n},$$

for suitable x_1, \dots, x_n in X . Then the function $f = f_{x_1} + \dots + f_{x_n}$ belongs to \mathcal{L} and $f > 1$ on X , therefore $1 = f \cap 1 \in \mathcal{L}$ by the condition 3°. \diamond

6.9.7. Lemma. If \mathcal{L} is a closed linear subspace of $\mathcal{C}_{\mathbb{R}}(X)$ satisfying 1°-3°, then \mathcal{L} is 2-fold transitive on X in the following sense:

$$\left. \begin{array}{l} x, y \in X \\ x \neq y \\ a, b \in \mathbb{R} \end{array} \right\} \Rightarrow \exists g \in \mathcal{L} \ni g(x) = a \text{ and } g(y) = b.$$

That is, for every pair of distinct points of X , there is a function in \mathcal{L} that takes on any specified values at the points.

Proof. Let $x, y \in X$, $x \neq y$, and let $a, b \in \mathbb{R}$. Choose (by 1°) a function $h \in \mathcal{L}$ such that $h(x) \neq h(y)$ and let $k = h - h(y)1$. Then $k \in \mathcal{L}$ (by the preceding lemma) and

$$k(y) = 0, \quad k(x) = h(x) - h(y) \neq 0.$$

Let $g_1 = (1/k(x))k$; then $g_1 \in \mathcal{L}$ and

$$g_1(x) = 1 \quad \text{and} \quad g_1(y) = 0.$$

Similarly, there exists a function $g_2 \in \mathcal{L}$ such that

$$g_2(x) = 0 \quad \text{and} \quad g_2(y) = 1,$$

and the function $g = ag_1 + bg_2$ in \mathcal{L} has the desired values $g(x) = a$ and $g(y) = b$. \diamond

Proof of Theorem 6.9.1: Let \mathcal{L} be as in the statement of the theorem. In view of Lemma 6.9.3, we can suppose that \mathcal{L} is a closed linear subspace of $\mathcal{C}_{\mathbb{R}}(X)$ satisfying 1°-3° and our problem is to show that $\mathcal{L} = \mathcal{C}_{\mathbb{R}}(X)$.

Given $f \in \mathcal{C}_{\mathbb{R}}(X)$ and $\epsilon > 0$, it will suffice to show that there is a function $g \in \mathcal{L}$ with $\|g - f\|_{\infty} \leq \epsilon$ (this will show that $f \in \overline{\mathcal{L}} = \mathcal{L}$).

The proof rests on the following two properties of \mathcal{L} (verified in Lemmas 6.9.5 and 6.9.7):

$$(A) \quad u, v \in \mathcal{L} \Rightarrow u \cup v, u \cap v \in \mathcal{L}.$$

$$(B) \quad \left. \begin{array}{l} x, y \in X \\ x \neq y \\ a, b \in \mathbb{R} \end{array} \right\} \Rightarrow \exists u \in \mathcal{L} \ni u(x) = a \text{ and } u(y) = b.$$

(In words, \mathcal{L} contains finite sups and infs, and is 2-fold transitive on X .)

For each pair of points $x, y \in X$, choose a function $g_{xy} \in \mathcal{L}$ such that

$$g_{xy}(x) = f(x), \quad g_{xy}(y) = f(y).$$

{If $x \neq y$, cite (B) with $a = f(x)$, $b = f(y)$; if $x = y$, let $g_{xx} = f(x)1$, which belongs to \mathcal{L} by Lemma 6.9.6.} Trivially,

$$(*) \quad |g_{xy}(x) - f(x)| < \epsilon \quad \text{and} \quad |g_{xy}(y) - f(y)| < \epsilon.$$

Let

$$U_{xy} = \{z \in X : g_{xy}(z) < f(z) + \epsilon\} = (g_{xy} - f)^{-1}((-\infty, \epsilon))$$

$$V_{xy} = \{z \in X : g_{xy}(z) > f(z) - \epsilon\} = (g_{xy} - f)^{-1}((-\epsilon, +\infty)).$$

By the continuity of $g_{xy} - f$, the sets U_{xy} , V_{xy} are open, and by (*) we have

$$(**) \quad x, y \in U_{xy} \quad \text{and} \quad x, y \in V_{xy}.$$

Fix a point $y \in X$. Construct a function g_y in \mathcal{L} as follows. The sets $(U_{xy})_{x \in X}$ form an open covering of X ; by compactness,

$$X = U_{x_1 y} \cup \dots \cup U_{x_n y}$$

for suitable points x_1, \dots, x_n (more precisely, $n = n(y)$ depends on y). The function

$$g_y = g_{x_1 y} \cap \dots \cap g_{x_n y}$$

belongs to \mathcal{L} by (A).

claim 1: $g_y < f + \epsilon 1$ on X .

Let $z \in X$. Say $z \in U_{x_i y}$. Then

$$g_y(z) \leq g_{x_i y}(z) < f(z) + \epsilon$$

(the first inequality by the definition of g_y , the second by the definition of $U_{x_i y}$), whence the claim.

Define

$$V_y = V_{x_1 y} \cap \dots \cap V_{x_n y};$$

the set V_y is open, and $y \in V_y$ by (**).

claim 2: $g_y > f - \epsilon 1$ on V_y .

Let $z \in V_y$. For all i , $z \in V_{x_i y}$ hence $g_{x_i y}(z) > f(z) - \epsilon$ (by the definition of $V_{x_i y}$); therefore $g_y(z) > f(z) - \epsilon$ (by the definition of g_y), whence the claim.

The family $(V_y)_{y \in X}$ is an open covering of the compact space X , so

$$X = V_{y_1} \cup \dots \cup V_{y_m}$$

for suitable points y_1, \dots, y_m in X . The function

$$g = g_{y_1} \cup \dots \cup g_{y_m}$$

belongs to \mathcal{L} by (A), and

$$g < f + \epsilon 1 \quad \text{on } X$$

by claim 1.

claim 3: $g > f - \epsilon 1$ on X .

Let $z \in X$. Say $z \in V_{y_j}$. Then

$$g(z) \geq g_{y_j}(z) > f(z) - \epsilon$$

(the first inequality by the definition of g , the second by claim 2), whence the claim.

Thus $f - \epsilon 1 < g < f + \epsilon 1$ on X , therefore $\|f - g\|_\infty \leq \epsilon$ (in fact, the inequality is strict, since the range of $f - g$ is a compact subset of \mathbb{R}). \diamond

The motivation for what follows is the observation that

$$|t| = \sqrt{t^2} \quad (t \in \mathbb{R});$$

the form of the right-hand side shows that, to approximate the function $t \mapsto |t|$ by polynomials in t , one need only approximate the square-root function by polynomial functions.

6.9.8. Lemma. In $C_{\mathbb{R}}[0, 1]$, the function $t \mapsto \sqrt{t}$ is the uniform limit of a sequence of polynomial functions without constant term.

Proof. Define a sequence of polynomial functions p_0, p_1, p_2, \dots recursively, as follows: $p_0(t) \equiv 0$ and

$$(*) \quad p_{n+1}(t) = p_n(t) + \frac{1}{2} [t - (p_n(t))^2].$$

It is clear (by induction) that the p_n are all polynomial functions and that $p_n(0) = 0$ for all n .

claim: $0 \leq p_0(t) \leq p_1(t) \leq \dots \leq p_n(t) \leq \sqrt{t}$ on $[0, 1]$.

The proof is by induction on n . For $n = 0$ the assertion is trivial. Assuming all's well for n , for every $t \in [0, 1]$ we have

$$(p_n(t))^2 \leq t, \quad \frac{1}{2} [t - (p_n(t))^2] \geq 0,$$

therefore $p_{n+1}(t) \geq p_n(t)$ by (*). Also

$$\begin{aligned}
 \sqrt{t} - p_{n+1}(t) &= \sqrt{t} - p_n(t) - \frac{1}{2}[t - (p_n(t))^2] \\
 (**) \quad &= [\sqrt{t} - p_n(t)] - \frac{1}{2}[\sqrt{t} - p_n(t)][\sqrt{t} + p_n(t)] \\
 &= [\sqrt{t} - p_n(t)]\{1 - \frac{1}{2}[\sqrt{t} + p_n(t)]\};
 \end{aligned}$$

but $\sqrt{t} + p_n(t) \leq \sqrt{t} + \sqrt{t} \leq 2$, therefore $\frac{1}{2}[\sqrt{t} + p_n(t)] \leq 1$; it follows that both factors in the rightmost member of (**) are ≥ 0 (the first factor, by the induction hypothesis), consequently $\sqrt{t} - p_{n+1}(t) \geq 0$, which completes the induction.

Define $f(t) = \sup_n p_n(t)$ for all $t \in [0, 1]$; thus

$$0 \leq f(t) \leq 1 \quad \text{and} \quad p_n(t) \uparrow f(t)$$

for all $t \in [0, 1]$. Passing to the limit in (*), we have

$$f(t) = f(t) + \frac{1}{2}[t - (f(t))^2],$$

whence $t - (f(t))^2 = 0$, $f(t) = \sqrt{t}$. Thus $p_n(t) \uparrow \sqrt{t}$ for all $t \in [0, 1]$. Since the p_n and the square-root function are continuous, it follows from Dini's theorem (6.2.24) that the p_n converge to the square-root function uniformly on $[0, 1]$. \diamond

6.9.9. Theorem. (Stone-Weierstrass theorem, real case) *Let \mathcal{A} be a subalgebra of $C_{\mathbb{R}}(X)$ such that*

1° \mathcal{A} separates the points of X , and

2° \mathcal{A} annihilates no point of X .

Then \mathcal{A} is uniformly dense in $C_{\mathbb{R}}(X)$.

Proof. By assumption, \mathcal{A} is closed under pointwise sums, products and scalar multiples, in particular, \mathcal{A} is a linear subspace of $C_{\mathbb{R}}$. Let

$$\mathcal{B} = \overline{\mathcal{A}}$$

be the uniform closure of \mathcal{A} in $C_{\mathbb{R}}$ (i.e., the closure of \mathcal{A} in $C_{\mathbb{R}}$ for the topology derived from the sup-metric). Since $\mathcal{B} \supset \mathcal{A}$ it is clear that \mathcal{B} also satisfies 1° and 2°.

claim: \mathcal{B} is a subalgebra of $C_{\mathbb{R}}$.

By the argument in Lemma 6.9.3, \mathcal{B} is a linear subspace of $C_{\mathbb{R}}$. If $f, g \in \mathcal{B}$ and if $(f_n), (g_n)$ are sequences in \mathcal{A} such that

$$f_n \rightarrow f, \quad g_n \rightarrow g \quad \text{uniformly,}$$

then $f_n g_n \rightarrow fg$ uniformly, as one sees from the computation

$$\begin{aligned}
 f_n g_n - fg &= (f_n - f)(g_n - g) + f(g_n - g) + (f_n - f)g, \\
 \|f_n g_n - fg\|_{\infty} &\leq \|f_n - f\|_{\infty} \|g_n - g\|_{\infty} + \|f\|_{\infty} \|g_n - g\|_{\infty} \\
 &\quad + \|f_n - f\|_{\infty} \|g\|_{\infty},
 \end{aligned}$$

therefore $fg \in \overline{\mathcal{A}} = \mathcal{B}$.

Changing notation, we can suppose that \mathcal{A} is closed for the uniform topology and our problem is to show that $\mathcal{A} = \mathcal{C}_{\mathbb{R}}$.

claim: If $f \in \mathcal{A}$ and $f \geq 0$ then $\sqrt{f} \in \mathcal{A}$.

By \sqrt{f} we mean the (continuous) function $x \mapsto \sqrt{f(x)}$ ($x \in X$). Passing to a scalar multiple of f , we can suppose that $0 \leq f \leq 1$. Let $s : [0, 1] \rightarrow \mathbb{R}$ be the usual square-root function $s(t) = \sqrt{t}$ ($0 \leq t \leq 1$), so that $\sqrt{f} = s \circ f$. By Lemma 6.9.8, there exists a sequence (p_n) of real polynomial functions, without constant term, such that $p_n \rightarrow s$ uniformly on $[0, 1]$. Since p_n has no constant term, $p_n \circ f \in \mathcal{A}$ (for example, if $p_n(t) \equiv a_1 t + a_2 t^2 + \dots + a_N t^N$ then $p_n \circ f = a_1 f + a_2 f^2 + \dots + a_N f^N \in \mathcal{A}$). Moreover, $p_n \circ f \rightarrow s \circ f$ uniformly on X , as one sees from the computation

$$\|p_n \circ f - s \circ f\|_{\infty} = \|(p_n - s) \circ f\|_{\infty} \leq \|p_n - s\|_{\infty}$$

(the first two sup-norms are calculated as x varies over X , the third as t varies over $[0, 1]$), consequently $s \circ f \in \overline{\mathcal{A}} = \mathcal{A}$, whence the claim. (Alternatively, $p_n \circ f \rightarrow s \circ f$ uniformly on X because $p_n \rightarrow s$ uniformly on $f(X) \subset [0, 1]$.)

It follows from the preceding claim, and the formula $|f| = \sqrt{f^2}$, that $f \in \mathcal{A} \Rightarrow |f| \in \mathcal{A}$. Summarizing: \mathcal{A} is a closed linear subspace of $\mathcal{C}_{\mathbb{R}}$, satisfying the conditions 1° and 2°, such that $f \in \mathcal{A} \Rightarrow |f| \in \mathcal{A}$ (hence \mathcal{A} is also closed under finite sups and infs). To complete the proof that $\mathcal{A} = \mathcal{C}_{\mathbb{R}}$, we need only show that \mathcal{A} also satisfies condition 3° of Theorem 6.9.1; since $f, g \in \mathcal{A} \Rightarrow f \cap g \in \mathcal{A}$, it will suffice to show that $1 \in \mathcal{A}$.

claim: $1 \in \mathcal{A}$.

For each $x \in X$ there exists (by condition 2°) a function $f_x \in \mathcal{A}$ with $f_x(x) \neq 0$. Replacing f_x by $(f_x)^2$ (or by $|f_x|$) we can suppose that

$$f_x \geq 0 \text{ on } X \text{ and } f_x(x) > 0.$$

Multiplying by a scalar, we can suppose further that

$$f_x(x) > 1.$$

The set

$$U_x = \{y \in X : f_x(y) > 1\}$$

is open and $x \in U_x$, thus the family $(U_x)_{x \in X}$ is an open covering of X ; by compactness,

$$X = U_{x_1} \cup \dots \cup U_{x_n}$$

for suitable points x_1, \dots, x_n . Define

$$f = f_{x_1} + \dots + f_{x_n};$$

then $f \in \mathcal{A}$ and $f > 1$ on X .

Note that if $r \in \mathbb{R}$, $r > 1$, then $r^{1/n} \downarrow 1$ (because $\log r > 0$, therefore $\log r^{1/n} = \frac{1}{n} \log r \downarrow 0 = \log 1$). It follows that $f^{1/n} \downarrow 1$ pointwise on X ,

therefore

$$f^{1/n} \downarrow 1 \quad \text{uniformly on } X$$

by Dini's theorem (6.2.24). In particular, the subsequence $f^{1/2^n}$ converges to 1 uniformly; since $f^{1/2^n} \in \mathcal{A}$ (by induction: $f^{1/2} \in \mathcal{A}$ and $f^{1/2^{n+1}} = \sqrt{f^{1/2^n}}$), it follows that $1 \in \overline{\mathcal{A}} = \mathcal{A}$ as claimed, and the proof of the theorem is complete (by the discussion preceding the claim). \diamond

6.9.10. Corollary. (Weierstrass approximation theorem) *For every continuous real-valued function $f : [a, b] \rightarrow \mathbb{R}$ on a closed interval $[a, b]$, there exists a sequence of real polynomial functions (p_n) such that $p_n \rightarrow f$ uniformly on $[a, b]$.*

Proof. A real polynomial function on $[a, b]$ is a function $p : [a, b] \rightarrow \mathbb{R}$ such that $p(t) = \sum_{k=0}^n c_k t^k$ for all $t \in [a, b]$, where c_0, c_1, \dots, c_n are suitable real numbers. Such functions are obviously continuous, and the set \mathcal{A} of all such functions is a subalgebra of $\mathcal{C}_{\mathbb{R}}[a, b]$ that meets the requirements of the preceding theorem: for example, the monomial function $p(t) \equiv t$ single-handedly separates all pairs of points of $[a, b]$, and the constant function $p(t) \equiv 1$ annihilates no point of $[a, b]$. \diamond

For continuous complex-valued functions, it is necessary to assume that the subalgebra is closed under complex-conjugation of functions:

6.9.11. Corollary. (Stone-Weierstrass theorem, complex case) *Let X be a compact space, \mathcal{B} a (complex) subalgebra of $\mathcal{C}_{\mathbb{C}}(X)$ such that*

- (i) \mathcal{B} separates the points of X ,
- (ii) \mathcal{B} annihilates no points of X , and
- (iii) $f \in \mathcal{B} \Rightarrow \overline{f} \in \mathcal{B}$ (where \overline{f} is the complex-conjugate of f).

Then \mathcal{B} is uniformly dense in $\mathcal{C}_{\mathbb{C}}(X)$.

Proof. Recall that $\overline{f}(x) = \overline{f(x)}$ for all $x \in X$, where $\overline{f(x)}$ is the conjugate of the complex number $f(x)$. As in the preceding section (cf. 6.8.11), we regard $\mathcal{C}_{\mathbb{R}}(X)$ as the \mathbb{R} -subalgebra of $\mathcal{C}_{\mathbb{C}}(X)$ consisting of all functions $f \in \mathcal{C}_{\mathbb{C}}(X)$ that are real-valued. Let

$$\mathcal{A} = \mathcal{B} \cap \mathcal{C}_{\mathbb{R}}(X)$$

be the set of real-valued functions in \mathcal{B} ; since both \mathcal{B} and $\mathcal{C}_{\mathbb{R}}(X)$ are \mathbb{R} -subalgebras of $\mathcal{C}_{\mathbb{C}}(X)$ (that is, subrings of $\mathcal{C}_{\mathbb{C}}(X)$ that are also \mathbb{R} -linear subspaces of $\mathcal{C}_{\mathbb{C}}(X)$), the same is true of \mathcal{A} . Moreover, it is clear from the assumption (iii) and the formulas

$$f = \operatorname{Re} f + i \operatorname{Im} f, \quad \operatorname{Re} f = \frac{1}{2}(f + \overline{f}), \quad \operatorname{Im} f = \frac{1}{2i}(f - \overline{f}),$$

valid for every $f \in \mathcal{C}_{\mathbb{C}}(X)$, that

$$(*) \quad f \in \mathcal{B} \quad \Leftrightarrow \quad \operatorname{Re} f, \operatorname{Im} f \in \mathcal{A}.$$

The idea of the proof is to apply the real case of the Stone–Weierstrass theorem to the subalgebra \mathcal{A} of $C_{\mathbb{R}}(X)$; to this end, let us verify that \mathcal{A} satisfies the conditions 1° and 2° of Theorem 6.9.9.

(1°) If $x, y \in X$, $x \neq y$, choose $f \in \mathcal{B}$ so that $f(x) \neq f(y)$; then one of $\operatorname{Re} f$, $\operatorname{Im} f$ is a function $g \in \mathcal{A}$ such that $g(x) \neq g(y)$.

(2°) If $x \in X$, choose $f \in \mathcal{B}$ so that $f(x) \neq 0$; then one of $\operatorname{Re} f$, $\operatorname{Im} f$ is a function $g \in \mathcal{A}$ such that $g(x) \neq 0$.

It now follows from Theorem 6.9.9 that \mathcal{A} is uniformly dense in $C_{\mathbb{R}}(X)$. To complete the proof, we need only show that every $f \in C_{\mathbb{C}}(X)$ is the uniform limit of a sequence of functions $f_n \in \mathcal{B}$. Write $f = g + ih$ with $g, h \in C_{\mathbb{R}}(X)$ (the real and imaginary parts of f), and let (g_n) , (h_n) be sequences in \mathcal{A} such that $g_n \rightarrow g$ and $h_n \rightarrow h$ uniformly. Then the functions $f_n = g_n + ih_n$ belong to \mathcal{B} by (*), and $f_n \rightarrow g + ih = f$ uniformly. \diamond

Exercises

1. In Theorem 6.9.1, none of the conditions 1°–3° can be omitted.

{Hint: Let $X = \{1, 2\} \subset \mathbb{R}$ be the discrete space with two points, so that $C_{\mathbb{R}}(X) = \mathcal{F}(X, \mathbb{R})$ can be identified with the set of all ordered pairs (x_1, x_2) of real numbers, that is, with \mathbb{R}^2 . Consider, in turn, the following linear subspaces of \mathbb{R}^2 : $\mathcal{L} = \{(c, c) : c \in \mathbb{R}\}$, the set of all constant functions; $\mathcal{L} = \{(0, c) : c \in \mathbb{R}\}$, the set of all functions that vanish at 1; $\mathcal{L} = \{(c, 2c) : c \in \mathbb{R}\}$, the set of all scalar multiples of the insertion mapping $\iota : X \rightarrow \mathbb{R}$, $\iota(x) = x$.}

2. Let K be a nonempty compact subset of \mathbb{C} , $u : K \rightarrow \mathbb{C}$ the insertion mapping $u(z) = z$ ($z \in K$), \bar{u} the conjugate function $\bar{u}(z) = \bar{z}$ ($z \in K$), and let \mathcal{B} be the subalgebra of $C_{\mathbb{C}}(K)$ generated by u and \bar{u} ; thus, a typical element of \mathcal{B} is a linear combination of functions of the form

$$z \mapsto z^m \bar{z}^n \quad (z \in K),$$

where m, n are nonnegative integers. The functions $p \in \mathcal{B}$ are called *polynomials in z and \bar{z}* , a typical such function having the form

$$p(z) = \sum_{m,n=0}^{\infty} c_{m,n} z^m \bar{z}^n \quad (z \in K),$$

where all but finitely many of the coefficients $c_{m,n}$ are equal to 0. (The underlying algebraic concept: the algebra $\mathbb{C}[s, t]$ of polynomials in two commuting indeterminates s and t ; such a polynomial determines a function of $z \in K$ via the substitutions $s \mapsto z$, $t \mapsto \bar{z}$.)

(i) \mathcal{B} is uniformly dense in $C_{\mathbb{C}}(K)$.

(ii) If $K = \mathbb{U} = \{z \in \mathbb{C} : |z| = 1\}$ (the unit circle in the complex plane)

then $\bar{z} = z^{-1}$ for all $z \in K$, and every $p \in \mathcal{B}$ can be written in the form

$$p(z) = \sum_{k \in \mathbb{Z}} c_k z^k \quad (z \in \mathbb{U})$$

where, in the notation of the earlier representation of p ,

$$c_k = \sum_{m-n=k} c_{m,n}$$

($= 0$ for all but finitely many integers k). If $p \in \mathcal{B}$ then the function $F : \mathbb{R} \rightarrow \mathbb{C}$ defined by $F(x) = p(e^{2\pi i x})$ is continuous, periodic of period 1, and has the representation

$$F(x) = \sum_{k \in \mathbb{Z}} c_k e^{2\pi i k x} = \sum_{k \in \mathbb{Z}} c_k [\cos 2\pi k x + i \sin 2\pi k x]$$

(such functions are called *trigonometric polynomials*).

(iii) With \mathbb{U} as in (ii), every continuous periodic function $F : \mathbb{R} \rightarrow \mathbb{C}$ of period 1 has a representation

$$F(x) = g(e^{2\pi i x}) \quad (x \in \mathbb{R})$$

for a suitable function $g \in \mathcal{C}_{\mathbb{C}}(\mathbb{U})$, hence is the uniform limit of trigonometric polynomials. The proof of the existence of g entails a slight digression into “quotient topologies”.⁴

3. A topological space is said to be *locally compact* if it is separated (6.1.6) and if each point of the space has a compact neighborhood (in which case *every* neighborhood of a point contains a compact neighborhood of the point).⁵ Let X be a noncompact, locally compact space (for example, $X = \mathbb{R}^n$ with the usual topology).

(i) A continuous function $f : X \rightarrow \mathbb{R}$ (or \mathbb{C}) is said to *vanish at infinity* if, for every $\epsilon > 0$, the set $K_{\epsilon} = \{x \in X : |f(x)| \geq \epsilon\}$ is compact. Such a function is necessarily bounded ($f(K_1)$ is compact and $|f(x)| < 1$ on $X - K_1$). The set $\mathcal{C}_0(X)$ of all continuous functions vanishing at infinity is a Banach algebra for the pointwise operations and the sup-norm $\|f\|_{\infty} = \sup_{x \in X} |f(x)|$ (cf. 6.8.14).

(iii) The Stone-Weierstrass theorem extends to X , provided that $\mathcal{C}(X)$ is replaced by $\mathcal{C}_0(X)$; that is, if \mathcal{A} is a subalgebra of $\mathcal{C}_0(X)$ that separates the points of X , annihilates no point of X , and is closed under complex conjugation, then \mathcal{A} is dense in $\mathcal{C}_0(X)$ for the norm topology.⁶

⁴ Cf. J. Dixmier, *General topology* [Springer-Verlag, New York, 1984], p. 83, Corollary 7.5.6.

⁵ *Op. cit.*, p. 46, Definition 4.5.2.

⁶ *Op. cit.*, p. 84, Corollary 7.5.8.

CHAPTER 7

Product Measure

§7.1. Extension of measures

§7.2. Product measures

§7.3. Iterated integrals, the case of finite measures

§7.4. Fubini–Tonelli theorem for σ -finite measures

If (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) are measure spaces, how can μ and ν be combined to define a measure on a suitable σ -algebra of subsets of $X \times Y$? The question is not academic: a satisfactory answer would open the door to constructing measures on $\mathbb{R}^2, \mathbb{R}^3, \dots$, starting with Lebesgue measure on \mathbb{R} . Lebesgue measure λ on \mathbb{R} assigns to an interval $[a, b]$ its length $b - a$; we expect the measure π on \mathbb{R}^2 derived from λ to assign to a rectangle $[a, b] \times [c, d]$ its area $(b - a)(d - c)$. Thus, writing $I = [a, b]$ and $J = [c, d]$, our expectation is that the ‘planar measure’ π on \mathbb{R}^2 derived from λ should satisfy the formula

$$\pi(I \times J) = \lambda(I)\lambda(J).$$

More generally, we would like to define a measure π on a suitable σ -algebra of subsets of $X \times Y$ that includes all ‘measurable rectangles’ $E \times F$ ($E \in \mathcal{S}$, $F \in \mathcal{T}$), such that the ‘product rule’

$$\pi(E \times F) = \mu(E)\nu(F)$$

holds for every measurable rectangle. Such a measure exists¹ but it is not necessarily unique (Exercise 2). However, when the given measures μ and ν are σ -finite, π is uniquely determined by the above ‘product formula’; inasmuch as the uniqueness property is highly desirable (and σ -finiteness is needed for other reasons as well), we shall limit the construction of π to the case of σ -finite measures. The first step is to define a set function on the algebra of sets generated by the measurable rectangles $E \times F$ and satisfying the product rule, then extend it to the σ -algebra generated

¹Cf. H.L. Royden, *Real analysis* [3rd edn., Macmillan, New York, 1988], p. 303ff; H.S. Bear, *A primer of Lebesgue integration* [Academic Press, New York, 1995], Chapter 14.

by the measurable rectangles; accordingly, we begin by considering a ‘measure’ defined on an algebra \mathcal{A} of sets, and show how to extend it to a measure on the σ -algebra $\mathcal{S}(\mathcal{A})$ generated by \mathcal{A} (2.4.4).

7.1. Extension of Measures

Recall (2.4.1) that an *algebra* of subsets of a set X is a nonempty set $\mathcal{A} \subset \mathcal{P}(X)$ that is closed under complementation and finite unions (thus a σ -algebra is an algebra that is closed under countable unions). In Chapter 2, measures were defined to have domain a σ -algebra (2.4.12); it is useful to allow the domain to be merely an algebra, but a ‘measure’ is still required to be countably additive in the appropriate sense:

7.1.1. Definition. Let X be a set, $\mathcal{A} \subset \mathcal{P}(X)$ an algebra of subsets of X . A function $\mu : \mathcal{A} \rightarrow [0, +\infty]$ is said to be a **measure** if

- 1° $\mu(\emptyset) = 0$, and
- 2° for every sequence (E_n) of pairwise disjoint sets in \mathcal{A} whose union $\bigcup E_n$ belongs to \mathcal{A} ,

$$\mu\left(\bigcup E_n\right) = \sum \mu(E_n).$$

So to speak, μ is countably additive ‘conditionally’, that is, provided that the countable unions in question belong to its domain.

7.1.2. Remark. Every measure μ is monotone: if $E, F \in \mathcal{A}$ with $E \subset F$, then $\mu(E) \leq \mu(F)$, as one sees on applying μ to the equality $F = E \cup (F - E) \cup \emptyset \cup \emptyset \cup \dots$.

Recall that an *outer measure* on X is a function $\rho : \mathcal{P}(X) \rightarrow [0, +\infty]$ that vanishes at \emptyset , is monotone, and is countably subadditive (2.2.7). Although measures defined on algebras are acceptable (inevitable), our first impulse is to extend them to σ -algebras, which have the advantage that they accommodate pointwise sequential limits (4.1.20). The basic strategy is straightforward: starting with a measure μ on an algebra \mathcal{A} , extend μ to an outer measure μ^* , then drop back to the σ -algebra $\mathcal{M}(\mu^*)$ of μ^* -measurable sets (2.2.7)—and note that this σ -algebra contains \mathcal{A} . Here’s how to get the outer measure (cf. 2.1.4):

7.1.3. Theorem. Let \mathcal{A} be an algebra of subsets of a set X , μ a measure on \mathcal{A} . For every subset A of X , define

$$\mu^*(A) = \inf\left\{\sum_{n=1}^{\infty} \mu(E_n) : E_n \in \mathcal{A} \ (n = 1, 2, 3, \dots), A \subset \bigcup_{n=1}^{\infty} E_n\right\}.$$

Then μ^* is an outer measure on X , and $\mu^*(E) = \mu(E)$ for all $E \in \mathcal{A}$.

Proof. Note that $A \subset X \in \mathcal{A}$, thus the displayed infimum is taken over a nonempty set.

Observe first that μ is *countably subadditive* on \mathcal{A} in the following sense: If (E_n) is a sequence in \mathcal{A} whose union $E = \bigcup E_n$ also belongs to \mathcal{A} , then

$$\mu(E) \leq \sum_{n=1}^{\infty} \mu(E_n).$$

For, defining $F_1 = E_1$ and $F_n = E_n - (E_1 \cup \dots \cup E_{n-1})$ for $n > 1$, the F_n are pairwise disjoint sets in \mathcal{A} with union $E \in \mathcal{A}$, and $F_n \subset E_n$ for all n , thus

$$\mu(E) = \sum \mu(F_n) \leq \sum \mu(E_n)$$

by the countable additivity (7.1.1) and monotonicity (7.1.2) of μ .

claim: If $E \in \mathcal{A}$ then $\mu^*(E) = \mu(E)$; that is, the set function μ^* is an extension of the set function μ .

For, $E \subset E \in \mathcal{A}$ shows that $\mu^*(E) \leq \mu(E)$ by the definition of μ^* . On the other hand, if (E_n) is a sequence of sets in \mathcal{A} with $E \subset \bigcup_{n=1}^{\infty} E_n$, then $E = \bigcup_{n=1}^{\infty} E \cap E_n$ with $E \cap E_n \in \mathcal{A}$ for all n , therefore

$$\mu(E) \leq \sum \mu(E \cap E_n) \leq \sum \mu(E_n)$$

by the countable subadditivity and monotonicity of μ ; varying over all coverings of E by such sequences (E_n) , we conclude that $\mu(E) \leq \mu^*(E)$. Thus $\mu^*(E) = \mu(E)$ as claimed.

The proof that μ^* is an outer measure follows the format of the proof of Theorem 2.1.6. \diamond

7.1.4. *Definition.* Let ρ be an outer measure on a set X . Recall that a set $B \subset X$ is said to be ρ -measurable if it splits every subset A of X additively, in the sense that

$$\rho(A) = \rho(A \cap B) + \rho(A \cap B')$$

for all $A \subset X$ (2.2.7). The set of all ρ -measurable sets will be denoted $\mathcal{M}(\rho)$.

7.1.5. *Theorem.* If ρ is an outer measure on a set X , then the set $\mathcal{M}(\rho)$ of all ρ -measurable sets is a σ -algebra of subsets of X , and the restriction of ρ to $\mathcal{M}(\rho)$ is a measure.

Proof. It is obvious that the empty set is ρ -measurable and that $\mathcal{M}(\rho)$ is closed under complementation. As noted in 2.2.6, $\mathcal{M}(\rho)$ is closed under countable unions and the restriction of ρ to $\mathcal{M}(\rho)$ is countably additive (cf. 2.2.5). \diamond

A measure on an algebra can always be extended to a measure on a σ -algebra:

7.1.6. Theorem. (Hahn–Kolmogorov extension theorem)¹ Let \mathcal{A} be an algebra of subsets of a set X and let μ be a measure on \mathcal{A} (7.1.1). Then the set $\mathcal{M}(\mu^*)$ of all μ^* -measurable sets is a σ -algebra containing \mathcal{A} , and the restriction of μ^* to $\mathcal{M}(\mu^*)$ is a measure that extends μ .

Proof. We know that μ^* is an outer measure (7.1.3), that $\mathcal{M}(\mu^*)$ is a σ -algebra and that the restriction of μ^* to $\mathcal{M}(\mu^*)$ is a measure (7.1.5). From Theorem 7.1.3 we also know that $\mu^*|_{\mathcal{A}} = \mu$, thus it remains only to show that $\mathcal{A} \subset \mathcal{M}(\mu^*)$. Assuming $E \in \mathcal{A}$ and $A \subset X$, we need only show that

$$\mu^*(A) \geq \mu^*(A \cap E) + \mu^*(A \cap E')$$

(cf. 2.2.3, (i)). We can suppose that $\mu^*(A) < +\infty$; given any $\epsilon > 0$, it will then suffice to show that

$$(*) \quad \mu^*(A) + \epsilon \geq \mu^*(A \cap E) + \mu^*(A \cap E').$$

By the definition of μ^* , there exists a sequence (E_n) of sets in \mathcal{A} such that $A \subset \bigcup E_n$ and

$$\sum \mu(E_n) \leq \mu^*(A) + \epsilon.$$

Then $A \cap E \subset \bigcup E_n \cap E$, where $E_n \cap E \in \mathcal{A}$ for all n , therefore

$$(1) \quad \mu^*(A \cap E) \leq \sum \mu(E_n \cap E);$$

similarly,

$$(2) \quad \mu^*(A \cap E') \leq \sum \mu(E_n \cap E').$$

By the finite additivity of μ ,

$$\mu(E_n \cap E) + \mu(E_n \cap E') = \mu(E_n)$$

for all n , thus the addition of (1) and (2) yields

$$\mu^*(A \cap E) + \mu^*(A \cap E') \leq \sum \mu(E_n) \leq \mu^*(A) + \epsilon,$$

which verifies (*). \diamond

With notations as in the preceding theorem, since $\mathcal{M}(\mu^*)$ is a σ -algebra containing \mathcal{A} , it contains the σ -algebra $\mathcal{S}(\mathcal{A})$ generated by \mathcal{A} (2.4.4), thus

$$\mathcal{A} \subset \mathcal{S}(\mathcal{A}) \subset \mathcal{M}(\mu^*).$$

It follows that the restriction of μ^* to $\mathcal{S}(\mathcal{A})$ is a measure extending μ ; stated formally:

¹Hans Hahn (1879–1934), A.N. Kolmogorov (1903–1987).

7.1.7. Corollary. If μ is a measure on an algebra \mathcal{A} of subsets of a set X , then the restriction of μ^* (the outer measure induced by μ) to the σ -algebra $\mathcal{S}(\mathcal{A})$ generated by \mathcal{A} is a measure that extends μ .

If μ is a measure on an algebra \mathcal{A} , we know that there exists at least one measure on $\mathcal{S}(\mathcal{A})$ that extends μ (7.1.7), but there may be more than one (Exercise 1). That's the bad news. The good news is that if μ is σ -finite (in the sense to be defined shortly) then its extension to a measure on $\mathcal{S}(\mathcal{A})$ is *unique*; the heart of the matter is the following special case:

7.1.8. Lemma. Let \mathcal{A} be an algebra of subsets of a set X , $\mathcal{S} = \mathcal{S}(\mathcal{A})$ the σ -algebra generated by \mathcal{A} . If μ_1 and μ_2 are finite measures on \mathcal{S} such that $\mu_1 = \mu_2$ on \mathcal{A} , then $\mu_1 = \mu_2$ on \mathcal{S} .

Proof. The proof to be given works equally well for real and complex measures (6.5.1). Let

$$\mathcal{M} = \{E \in \mathcal{S} : \mu_1(E) = \mu_2(E)\}.$$

By assumption $\mathcal{A} \subset \mathcal{M}$, and our problem is to show that $\mathcal{S}(\mathcal{A}) \subset \mathcal{M}$; since \mathcal{A} is an algebra, it will suffice to show that \mathcal{M} is a monotone class (4.6.6). Note that if $E, F \in \mathcal{M}$ and $E \subset F$, then $F - E \in \mathcal{M}$ by the subtractivity of finite measures (2.6.1).

Assuming (E_n) is an increasing sequence of sets in \mathcal{M} with union E , let us show that $E \in \mathcal{M}$. At any rate $E \in \mathcal{S}$, and

$$E = E_1 \cup (E_2 - E_1) \cup (E_3 - E_2) \cup \dots;$$

since the terms of the union are pairwise disjoint and since, for $i = 1, 2$, μ_i is countably additive, we have

$$\begin{aligned} \mu_i(E) &= \mu_i(E_1) + \sum_{n=1}^{\infty} \mu_i(E_{n+1} - E_n) \\ &= \mu_i(E_1) + \lim_{n \rightarrow \infty} \sum_{k=1}^n \mu_i(E_{k+1} - E_k) \\ &= \mu_i(E_1) + \lim_{n \rightarrow \infty} \mu_i \left(\bigcup_{k=1}^n (E_{k+1} - E_k) \right) \\ &= \mu_i(E_1) + \lim_{n \rightarrow \infty} \mu_i(E_{n+1} - E_1) \\ &= \lim_{n \rightarrow \infty} [\mu_i(E_1) + \mu_i(E_{n+1} - E_1)] \\ &= \lim_{n \rightarrow \infty} \mu_i(E_{n+1}) = \lim_{n \rightarrow \infty} \mu_i(E_n); \end{aligned}$$

since $\mu_1(E_n) = \mu_2(E_n)$ for all n , we conclude that

$$\mu_1(E) = \lim_{n \rightarrow \infty} \mu_1(E_n) = \lim_{n \rightarrow \infty} \mu_2(E_n) = \mu_2(E),$$

thus $E \in \mathcal{M}$.

Finally, if (E_n) is a decreasing sequence in \mathcal{M} with intersection $\bigcap_{n=1}^{\infty} E_n = E \in \mathcal{S}$, then $E_1 - E_n \uparrow E_1 - E \in \mathcal{S}$; since $E_1 - E_n \in \mathcal{M}$ for all n , it follows from the preceding paragraph that $E_1 - E \in \mathcal{M}$, therefore $E = E_1 - (E_1 - E) \in \mathcal{M}$. Thus \mathcal{M} is a monotone class and the proof is complete. \diamond

The next lemma requires a new concept:

7.1.9. *Definition.* Let \mathcal{A} be an algebra of subsets of a set X . A measure μ on \mathcal{A} is said to be σ -finite if, for every $E \in \mathcal{A}$, there exists a sequence (E_n) of sets in \mathcal{A} such that $E = \bigcup_{n=1}^{\infty} E_n$ and $\mu(E_n) < +\infty$ for all n . {It is the same to say that X is the union of a sequence of sets in \mathcal{A} of finite measure.}

7.1.10. *Example.* (1) Lebesgue measure λ (2.2.9) is σ -finite, as is the restriction of λ to any algebra of subsets of \mathbb{R} that contains a sequence of bounded intervals with union \mathbb{R} .

(2) If X is a countable set, then the discrete measure on X is σ -finite (§2.4, Exercise 3, (i)).

(3) The trivial measure on a nonempty set is *not* σ -finite (§2.4, Exercise 3, (ii)).

7.1.11. *Lemma.* Let \mathcal{A} be an algebra of subsets of a set X , $\mathcal{S} = \mathcal{S}(\mathcal{A})$ the σ -algebra generated by \mathcal{A} . If μ_1 and μ_2 are measures on \mathcal{S} such that

- 1° $\mu_1 = \mu_2$ on \mathcal{A} , and
 - 2° the measure $\mu_1|_{\mathcal{A}} = \mu_2|_{\mathcal{A}}$ is σ -finite,
- then $\mu_1 = \mu_2$ on \mathcal{S} .

Proof. Let $E \in \mathcal{S}$; we are to show that $\mu_1(E) = \mu_2(E)$. By the hypothesis 2°, there exists a sequence (F_n) of sets in \mathcal{A} such that

$$X = \bigcup_{n=1}^{\infty} F_n, \quad \mu_i(F_n) < +\infty \text{ for all } n.$$

We can suppose that the F_n are pairwise disjoint (if they are not, they can be ‘disjointified’ as in the proof of 7.1.3). Then $E = \bigcup_{n=1}^{\infty} F_n \cap E$ is a disjoint union, therefore

$$\mu_i(E) = \sum_{n=1}^{\infty} \mu_i(F_n \cap E) \quad (i = 1, 2);$$

it will thus suffice to show that

$$\mu_1(F_n \cap E) = \mu_2(F_n \cap E) \quad \text{for all } n.$$

Thus, assuming $F \in \mathcal{A}$ and $\mu_i(F) < +\infty$, it will suffice to show that

$$(*) \quad \mu_1(F \cap E) = \mu_2(F \cap E) \quad \text{for all } E \in \mathcal{S}.$$

For $i = 1, 2$ define $\nu_i : \mathcal{S} \rightarrow \mathbb{R}$ by the formula

$$\nu_i(E) = \mu_i(F \cap E) \quad (\forall E \in \mathcal{S}).$$

It is clear that the ν_i are finite measures on \mathcal{S} , and $\nu_1 = \nu_2$ on \mathcal{A} by the hypothesis 1°, therefore $\nu_1 = \nu_2$ on \mathcal{S} by Lemma 7.1.8, thus (*) is verified. \diamond

The key extension theorem follows at once:

7.1.12. Theorem. (Unique extension theorem) *Let \mathcal{A} be an algebra of subsets of a set X , let μ be a σ -finite measure on \mathcal{A} (7.1.9), and let $\mathcal{S} = \mathcal{S}(\mathcal{A})$ be the σ -algebra generated by \mathcal{A} . Then there exists a unique measure $\bar{\mu}$ on \mathcal{S} that extends μ ; explicitly, $\bar{\mu} = \mu^*|_{\mathcal{S}}$.*

Proof. Existence: Corollary 7.1.7. Uniqueness: the preceding lemma. \diamond

Exercises

1. In contrast with the unique extension theorem, this exercise exhibits an algebra \mathcal{A} of sets such that the σ -algebra $\mathcal{S}(\mathcal{A})$ generated by \mathcal{A} admits two distinct measures that coincide on \mathcal{A} .²

Let \mathbb{Q} be the set of all rational numbers, regarded as an ordered set in the usual way. ‘Intervals’ are defined as in any ordered set having no smallest nor largest element; for example, if $a, b \in \mathbb{Q}$, $a < b$, then

$$\begin{aligned} [a, +\infty) &= \{r \in \mathbb{Q} : r \geq a\} \\ (-\infty, b) &= \{r \in \mathbb{Q} : r < b\} \\ [a, b] &= [a, +\infty) \cap (-\infty, b) = \{r \in \mathbb{Q} : a \leq r < b\}. \end{aligned}$$

Let \mathcal{C} be the set of subsets of \mathbb{Q} consisting of all of the three kinds of ‘intervals’ just described, together with the sets \emptyset and \mathbb{Q} .

(i) If $A, B \in \mathcal{C}$ then $A \cap B \in \mathcal{C}$; and $\mathbb{Q} - A$ either belongs to \mathcal{C} or is the union of two disjoint sets of \mathcal{C} .

(ii) The algebra $\mathcal{A} = \mathcal{A}(\mathcal{C})$ generated by \mathcal{C} consists of the set of all finite disjoint unions of sets in \mathcal{C} .

(iii) Every nonempty element of \mathcal{A} is infinite.

(iv) The σ -algebra generated by \mathcal{A} is the set of all subsets of \mathbb{Q} , that is, $\mathcal{S}(\mathcal{A}) = \mathcal{P}(\mathbb{Q})$.

(v) Let μ be the ‘counting measure’ on $\mathcal{S}(\mathcal{A}) = \mathcal{P}(\mathbb{Q})$, that is, for every subset E of \mathbb{Q} ,

$$\mu(E) = \begin{cases} n & \text{if } E \text{ is a finite set with } n \text{ elements} \\ +\infty & \text{if } E \text{ is infinite.} \end{cases}$$

²P.R. Halmos, *Measure theory* [Van Nostrand, New York, 1950; reprinted Springer-Verlag, New York, 1974], p. 57 (§13, Exercise 5).

Then μ and 2μ are distinct measures on $\mathcal{S}(\mathcal{A})$ such that $\mu(A) = (2\mu)(A)$ for all $A \in \mathcal{A}$.

{Hints: (ii) Cf. 4.6.1. (iv) $\{a\} = \bigcap_{n=1}^{\infty} [a, a + 1/n)$ for all $a \in \mathbb{Q}$. (v) μ is the ‘discrete measure’ of §2.4, Exercise 3.}

2. With notations as in Exercise 1, let

$$\mathcal{D} = \{A \times B : A, B \in \mathcal{A}\},$$

and let \mathcal{B} be the algebra of subsets of $\mathbb{Q} \times \mathbb{Q}$ generated by \mathcal{D} .

(i) If $D_1, D_2 \in \mathcal{D}$ then $D_1 \cap D_2 \in \mathcal{D}$; and if $D \in \mathcal{D}$ then $\mathbb{Q} \times \mathbb{Q} - D$ either belongs to \mathcal{D} or is the union of two disjoint sets of \mathcal{D} .

(ii) \mathcal{B} is the set of all finite disjoint unions of sets in \mathcal{D} .

(iii) Every nonempty element of \mathcal{D} (hence of \mathcal{B}) is infinite.

(iv) The σ -algebra generated by \mathcal{B} is the set $\mathcal{P}(\mathbb{Q} \times \mathbb{Q})$ of all subsets of $\mathbb{Q} \times \mathbb{Q}$.

(v) If π_1 is the counting measure on $\mathcal{P}(\mathbb{Q} \times \mathbb{Q})$ and if $\pi_2 = 2\pi_1$, then $\pi_1 \neq \pi_2$ but

$$\pi_1(A \times B) = \mu(A)\mu(B) = \pi_2(A \times B)$$

for all $A, B \in \mathcal{A}$, where μ is the measure on $\mathcal{P}(\mathbb{Q})$ described in Exercise 1 (indeed, $\pi_1 = \pi_2$ on \mathcal{B}).

3. Let \mathcal{A} be an algebra of subsets of a set X , $\mathcal{S} = \mathcal{S}(\mathcal{A})$ the σ -algebra generated by \mathcal{A} . If μ_1 and μ_2 are measures on \mathcal{S} such that

1° $\mu_1 \leq \mu_2$ on \mathcal{A} , and

2° $\mu_2|_{\mathcal{A}}$ is σ -finite (hence so is $\mu_1|_{\mathcal{A}}$, by 1°),

then $\mu_1 \leq \mu_2$ on \mathcal{S} .

{Hint: Adapt Lemma 7.1.8 to the case that $\mu_1 \leq \mu_2$ (employing Corollaries 2.6.2 and 2.6.3 instead of countable additivity), then modify the argument in Lemma 7.1.11.}

7.2. Product Measures

Notations fixed for the rest of the section: (X, \mathcal{S}) and (Y, \mathcal{T}) are measurable spaces (4.1.1).

Eventually we will contemplate σ -finite measures μ on \mathcal{S} and ν on \mathcal{T} , with a view towards defining a ‘product measure’ π on a suitable σ -algebra of subsets of $X \times Y$. The σ -algebra in question is the following:

7.2.1. Definition. The σ -algebra of subsets of $X \times Y$ generated by the sets $E \times F$ ($E \in \mathcal{S}$, $F \in \mathcal{T}$) is denoted $\mathcal{S} \times \mathcal{T}$ and is called the **product σ -algebra** of \mathcal{S} and \mathcal{T} (in that order).

{This is a slight abuse of notation, as the actual *product set* of \mathcal{S} and \mathcal{T} would consist of the ordered pairs (E, F) .}

In the context of the measurable space $(X \times Y, \mathcal{S} \times \mathcal{T})$, the 'measurable sets' are the sets that belong to $\mathcal{S} \times \mathcal{T}$. In particular, the sets $E \times F$ ($E \in \mathcal{S}$, $F \in \mathcal{T}$) are called **measurable rectangles**.

{CAUTION: If $A \subset X$ then $A \times \emptyset = \emptyset \in \mathcal{S} \times \mathcal{T}$, even if $A \notin \mathcal{S}$. Later on, we shall see that if $A \times B \in \mathcal{S} \times \mathcal{T}$ and $A \times B \neq \emptyset$, then $A \in \mathcal{S}$ and $B \in \mathcal{T}$ (7.2.7 below).}

7.2.2. Lemma. *The set \mathcal{A} of all finite disjoint unions of measurable rectangles $E \times F$ ($E \in \mathcal{S}$, $F \in \mathcal{T}$) is an algebra of subsets of $X \times Y$.*

Proof. Since \mathcal{S} and \mathcal{T} are closed under finite intersections and complementation, the lemma is immediate from Example 4.6.3. \diamond

7.2.3. Remark. Writing $\mathcal{P} = \{E \times F : E \in \mathcal{S}, F \in \mathcal{T}\}$, the algebra generated by \mathcal{P} is the algebra \mathcal{A} of the lemma, and $\mathcal{P} \subset \mathcal{A} \subset \mathcal{S} \times \mathcal{T}$; since $\mathcal{S} \times \mathcal{T}$ is the σ -algebra generated by \mathcal{P} , it is also the σ -algebra generated by \mathcal{A} .

The next definition prepares the way for analyzing sets $M \in \mathcal{S} \times \mathcal{T}$ in terms of sets in \mathcal{S} and \mathcal{T} ; to paraphrase the case that $X = Y = \mathbb{R}$, one slices M by lines parallel to the coordinate axes:

7.2.4. Definition. Let $M \subset X \times Y$. For each $x \in X$, we write

$$M_x = \{y \in Y : (x, y) \in M\},$$

called the **x -section** of M ; it is, so to speak, the result of intersecting M with the 'vertical line' $\{x\} \times Y$, then projecting the intersection onto the Y -axis:

$$M_x = \text{pr}_Y[M \cap (\{x\} \times Y)].$$

Alternatively, writing $i_x : Y \rightarrow X \times Y$ for the mapping $i_x(y) = (x, y)$,

$$M_x = i_x^{-1}(M).$$

Similarly, for each $y \in Y$, one defines the **y -section** of M , denoted M^y , by the formula

$$M^y = \{x \in X : (x, y) \in M\} = \text{pr}_X[M \cap (X \times \{y\})],$$

and, defining $i^y : X \rightarrow X \times Y$ by the formula $i^y(x) = (x, y)$, we have

$$M^y = (i^y)^{-1}(M).$$

7.2.5. Example. Suppose $M = A \times B$, where $A \subset X$ and $B \subset Y$. For $x \in X$,

$$(A \times B)_x = \begin{cases} B & \text{if } x \in A \\ \emptyset & \text{if } x \notin A \end{cases}$$

and for $y \in Y$,

$$(A \times B)^y = \begin{cases} A & \text{if } y \in B \\ \emptyset & \text{if } y \notin B. \end{cases}$$

For example,

$$(A \times B)_x = \text{pr}_Y[(A \times B) \cap (\{x\} \times Y)] = \text{pr}_Y[(A \cap \{x\}) \times B]$$

and $A \cap \{x\} = \{x\}$ or \emptyset according as $x \in A$ or $x \notin A$.

7.2.6. Lemma. *If $M \in \mathcal{S} \times \mathcal{T}$ then every section of M is a measurable set, that is, $M_x \in \mathcal{T}$ for all $x \in X$, and $M^y \in \mathcal{S}$ for all $y \in Y$.*

Proof. We give the proof for x -sections. Fix $x \in X$; to simplify the notation, we write $g = i_x$, thus

$$g: Y \rightarrow X \times Y, \quad g(y) = (x, y) \text{ for all } y \in Y.$$

As noted in Definition 7.2.4, $g^{-1}(M) = M_x$ for all $M \subset X \times Y$. Let

$$\mathcal{U} = \{M \subset X \times Y: M_x \in \mathcal{T}\} = \{M \subset X \times Y: g^{-1}(M) \in \mathcal{T}\}.$$

Since \mathcal{T} is a σ -algebra, it is clear from the formulas

$$g^{-1}(\emptyset) = \emptyset, \quad g^{-1}\left(\bigcup_{n=1}^{\infty} M_n\right) = \bigcup_{n=1}^{\infty} g^{-1}(M_n), \quad g^{-1}(X \times Y - M) = Y - g^{-1}(M)$$

that \mathcal{U} is a σ -algebra of subsets of $X \times Y$. If $M = E \times F$ ($E \in \mathcal{S}$, $F \in \mathcal{T}$) then $g^{-1}(M) = F$ or \emptyset according as $x \in E$ or $x \notin E$; in either case, $g^{-1}(M) \in \mathcal{T}$, so $M \in \mathcal{U}$. Thus \mathcal{U} is a σ -algebra containing all measurable rectangles, therefore it contains the σ -algebra $\mathcal{S} \times \mathcal{T}$ they generate; in other words, $g^{-1}(M) \in \mathcal{T}$ for all $M \in \mathcal{S} \times \mathcal{T}$. \diamond

7.2.7. Remark. If A and B are nonempty subsets of X and Y , respectively, such that $A \times B \in \mathcal{S} \times \mathcal{T}$, then $A \in \mathcal{S}$ and $B \in \mathcal{T}$. {For, choosing $a \in A$ and $b \in B$, we have $A = (A \times B)^b \in \mathcal{S}$ and $B = (A \times B)_a \in \mathcal{T}$ by the preceding lemma.}

We now contemplate measures on \mathcal{S} and \mathcal{T} ; henceforth:

$$(X, \mathcal{S}, \mu) \text{ and } (Y, \mathcal{T}, \nu) \text{ are measure spaces.}$$

7.2.8. Definition. For each set $M \in \mathcal{S} \times \mathcal{T}$, Lemma 7.2.6 permits us to define a function $f_M: X \rightarrow [0, +\infty]$ by the formula

$$f_M(x) = \nu(M_x) \quad (x \in X)$$

and a function $g^M: Y \rightarrow [0, +\infty]$ by the formula

$$g^M(y) = \mu(M^y) \quad (y \in Y).$$

7.2.9. *Remark.* In the notations of the definition, if ν is a finite measure, then the function f_M is bounded: $0 \leq f_M \leq \nu(Y) < +\infty$. Similarly, if μ is finite then $0 \leq g^M \leq \mu(X) < +\infty$.

7.2.10. *Lemma.* If ν is a finite measure on \mathcal{T} , then f_M is measurable with respect to \mathcal{S} for every $M \in \mathcal{S} \times \mathcal{T}$.

Proof. At any rate, the f_M are real-valued (and bounded) by the preceding remark. Let

$$\mathcal{U} = \{M \in \mathcal{S} \times \mathcal{T} : f_M \text{ is measurable with respect to } \mathcal{S}\};$$

the strategy of the proof is to show that \mathcal{U} is a monotone class containing the algebra \mathcal{A} generated by the measurable rectangles, hence containing the σ -algebra $\mathcal{S} \times \mathcal{T}$ generated by \mathcal{A} (whence $\mathcal{U} = \mathcal{S} \times \mathcal{T}$).

claim 1: \mathcal{U} contains every measurable rectangle.

Suppose $M = E \times F$ with $E \in \mathcal{S}$ and $F \in \mathcal{T}$. From Example 7.2.5 we see that

$$f_{E \times F}(x) = \nu((E \times F)_x) = \begin{cases} \nu(F) & \text{if } x \in E \\ 0 & \text{if } x \notin E; \end{cases}$$

thus $f_{E \times F} = \nu(F)\varphi_E$, where φ_E is the characteristic function of E , consequently $f_{E \times F}$ is measurable with respect to \mathcal{S} , that is, $E \times F \in \mathcal{U}$.

claim 2: \mathcal{U} is closed under finite disjoint unions.

Suppose $M, N \in \mathcal{U}$ and $M \cap N = \emptyset$. For each $x \in X$, writing $i_x(y) = (x, y)$ as in Definition 7.2.4, we have

$$\begin{aligned} (M \cup N)_x &= i_x^{-1}(M \cup N) = i_x^{-1}(M) \cup i_x^{-1}(N) = M_x \cup N_x, \\ M_x \cap N_x &= i_x^{-1}(M) \cap i_x^{-1}(N) = i_x^{-1}(M \cap N) = \emptyset, \end{aligned}$$

thus M_x and N_x are disjoint sets in \mathcal{T} and

$$\begin{aligned} f_{M \cup N}(x) &= \nu((M \cup N)_x) = \nu(M_x \cup N_x) \\ &= \nu(M_x) + \nu(N_x) = f_M(x) + f_N(x); \end{aligned}$$

thus $f_{M \cup N} = f_M + f_N$ is the sum of two measurable functions (because $M, N \in \mathcal{U}$) hence is measurable, that is, $M \cup N \in \mathcal{U}$.

claim 3: \mathcal{U} is closed under complementation.

Assuming $M \in \mathcal{U}$, we are to show that $M' = X \times Y - M$ belongs to \mathcal{U} . For every $x \in X$ we have

$$(M')_x = i_x^{-1}(X \times Y - M) = Y - i_x^{-1}(M) = Y - M_x,$$

therefore

$$f_{M'}(x) = \nu((M')_x) = \nu(Y - M_x) = \nu(Y) - \nu(M_x) = \nu(Y) - f_M(x),$$

thus $f_{M'} = \nu(Y) \cdot 1 - f_M$; by assumption, f_M is measurable with respect to \mathcal{S} , therefore so is $f_{M'}$, that is, $M' \in \mathcal{U}$.

claim 4: \mathcal{U} is a monotone class.

If (M_n) is an increasing sequence of sets in \mathcal{U} with union M then, for each $x \in X$, the sets $(M_n)_x = i_x^{-1}(M_n)$ form an increasing sequence of sets in \mathcal{T} and

$$\begin{aligned} M_x &= \left(\bigcup_{n=1}^{\infty} M_n\right)_x = i_x^{-1}\left(\bigcup_{n=1}^{\infty} M_n\right) \\ &= \bigcup_{n=1}^{\infty} i_x^{-1}(M_n) = \bigcup_{n=1}^{\infty} (M_n)_x, \end{aligned}$$

thus $(M_n)_x \uparrow M_x$ and

$$0 \leq f_{M_n}(x) = \nu((M_n)_x) \uparrow \nu(M_x) = f_M(x).$$

Briefly, $f_{M_n} \uparrow f_M$. Since the f_{M_n} are by assumption measurable, so is their (real-valued) pointwise limit f_M , that is, $M \in \mathcal{U}$.

Similarly, if $M_n \in \mathcal{U}$ and $M_n \downarrow M$, then $(M_n)_x \downarrow M_x$ for all $x \in X$, whence $f_{M_n}(x) \downarrow f_M(x)$ (because ν is finite); since the f_{M_n} are measurable, so is their pointwise limit f_M , that is, $M \in \mathcal{U}$. Thus \mathcal{U} is a monotone class. (Alternatively, one could cite the preceding paragraph and closure under complementation.)

From claims 1 and 2, we know that \mathcal{U} contains the algebra \mathcal{A} generated by the measurable rectangles (7.2.2), and \mathcal{U} is a monotone class by claim 4, therefore $\mathcal{S} \times \mathcal{T} \subset \mathcal{U}$ by the lemma on monotone classes (4.6.6). \diamond

7.2.11. Theorem. *If μ and ν are finite measures on the σ -algebras \mathcal{S} and \mathcal{T} , respectively, then there exists a unique measure π on $\mathcal{S} \times \mathcal{T}$ such that*

$$\pi(E \times F) = \mu(E)\nu(F) \quad \text{for all } E \in \mathcal{S} \text{ and } F \in \mathcal{T}$$

(in particular, π is finite). Explicitly, if $M \in \mathcal{S} \times \mathcal{T}$ then f_M is μ -integrable, g^M is ν -integrable, and

$$\pi(M) = \int f_M d\mu = \int g^M d\nu.$$

Proof. Let $M \in \mathcal{S} \times \mathcal{T}$. Since ν is finite, f_M is bounded (7.2.9) and measurable with respect to \mathcal{S} (7.2.10), therefore integrable with respect to the finite measure μ (4.4.20). We may therefore define a set function π on $\mathcal{S} \times \mathcal{T}$ by the formula

$$(*) \quad \pi(M) = \int f_M d\mu \quad (M \in \mathcal{S} \times \mathcal{T}).$$

Obviously $0 \leq \pi(M) < +\infty$ for all $M \in \mathcal{S} \times \mathcal{T}$.

If $M = E \times F$ with $E \in \mathcal{S}$ and $F \in \mathcal{T}$ then, as shown in “claim 1” of the preceding lemma, $f_{E \times F} = \nu(F)\varphi_E$, therefore

$$\pi(E \times F) = \int f_{E \times F} d\mu = \nu(F) \int \varphi_E d\mu = \mu(E)\nu(F).$$

Let us show that π is a (finite) measure.

We observe first that π is finitely additive. For, if $M, N \in \mathcal{S} \times \mathcal{T}$ with $M \cap N = \emptyset$ then, as shown in “claim 2” of the preceding lemma, $f_{M \cup N} = f_M + f_N$, therefore

$$\begin{aligned} \pi(M \cup N) &= \int f_{M \cup N} d\mu = \int (f_M + f_N) d\mu \\ &= \int f_M d\mu + \int f_N d\mu = \pi(M) + \pi(N). \end{aligned}$$

To complete the proof that π is a measure, we need only show that if (M_n) is an increasing sequence in $\mathcal{S} \times \mathcal{T}$ with union M , then $\pi(M_n) \uparrow \pi(M)$ (§2.6, Exercise 3). Indeed, as shown in the proof of “claim 4” of the preceding lemma, $f_{M_n} \uparrow f_M$ pointwise on X , therefore

$$\pi(M_n) = \int f_{M_n} d\mu \uparrow \int f_M d\mu = \pi(M)$$

by the monotone convergence theorem (4.5.3).

Summarizing, the formula $(*)$ defines a finite measure π on $\mathcal{S} \times \mathcal{T}$ such that $\pi(E \times F) = \mu(E)\nu(F)$ for all measurable rectangles. If π' is another measure on $\mathcal{S} \times \mathcal{T}$ such that $\pi'(E \times F) = \mu(E)\nu(F)$ for all measurable rectangles, then $\pi = \pi'$ on the set of all finite disjoint unions of measurable rectangles, that is, on the algebra \mathcal{A} generated by the measurable rectangles (7.2.2); in particular, $\pi'(X \times Y) = \mu(X)\nu(Y) < +\infty$ shows that π' is also a finite measure, consequently $\pi = \pi'$ on the σ -algebra $\mathcal{S} \times \mathcal{T}$ generated by \mathcal{A} , by the unique extension theorem (7.1.12).

Similarly, for every $M \in \mathcal{S} \times \mathcal{T}$, g^M is bounded (7.2.9) and measurable with respect to \mathcal{T} (by the obvious analogue of Lemma 7.2.10), hence integrable with respect to the finite measure ν , and the formula

$$\pi'(M) = \int g^M d\nu \quad (M \in \mathcal{S} \times \mathcal{T})$$

defines a measure π' on $\mathcal{S} \times \mathcal{T}$ such that $\pi'(E \times F) = \mu(E)\nu(F)$ for every measurable rectangle. Finally, $\pi' = \pi$ by the preceding paragraph, thus

$$\int f_M d\mu = \pi(M) = \pi'(M) = \int g^M d\nu$$

for all $M \in \mathcal{S} \times \mathcal{T}$. \diamond

The construction of product measure for σ -finite measures follows easily from the finite case:

7.2.12. Corollary. If μ and ν are σ -finite measures on the σ -algebras \mathcal{S} and \mathcal{T} , respectively, then there exists a unique measure π on $\mathcal{S} \times \mathcal{T}$ such that

$$\pi(E \times F) = \mu(E)\nu(F) \quad \text{for all } E \in \mathcal{S} \text{ and } F \in \mathcal{T}.$$

In particular, π is σ -finite.

Proof. By the σ -finiteness of μ , there exists a sequence (P_n) of sets in \mathcal{S} such that $X = \bigcup_{n=1}^{\infty} P_n$ and $\mu(P_n) < +\infty$ for all n ; replacing P_n by $P_1 \cup \dots \cup P_n$, we can suppose that $P_n \uparrow X$. Similarly, there exists a sequence (Q_n) in \mathcal{T} such that $Q_n \uparrow Y$ and $\nu(Q_n) < +\infty$ for all n .

Let \mathcal{A} be the set of all finite disjoint unions of measurable rectangles $E \times F$; thus \mathcal{A} is an algebra (7.2.2) and $\mathcal{S} \times \mathcal{T}$ is the σ -algebra generated by \mathcal{A} (7.2.3).

Existence: For each index n , define set functions $\mu_n : \mathcal{S} \rightarrow [0, +\infty]$ and $\nu_n : \mathcal{T} \rightarrow [0, +\infty]$ by the formulas

$$\begin{aligned} \mu_n(E) &= \mu(P_n \cap E) & (E \in \mathcal{S}), \\ \nu_n(F) &= \nu(Q_n \cap F) & (F \in \mathcal{T}). \end{aligned}$$

It is clear that the μ_n are finite measures on \mathcal{S} and that $\mu_n \uparrow \mu$ on \mathcal{S} ; similarly, the ν_n are finite measures on \mathcal{T} with $\nu_n \uparrow \nu$.

By the preceding theorem, for each n there exists a measure π_n on $\mathcal{S} \times \mathcal{T}$ such that

$$\pi_n(E \times F) = \mu_n(E)\nu_n(F) \quad \text{for all } E \in \mathcal{S}, F \in \mathcal{T};$$

since $\mu_n(E) \uparrow \mu(E)$ and $\nu_n(F) \uparrow \nu(F)$, it is easy to see that

$$(*) \quad \pi_n(E \times F) \uparrow \mu(E)\nu(F)$$

for all $E \in \mathcal{S}$ and $F \in \mathcal{T}$ (§1.15, Exercise 4, (ii)).¹

claim: $\pi_1 \leq \pi_2 \leq \pi_3 \leq \dots$

It is enough to show that $\pi_1 \leq \pi_2$. Since $\mu_1 \leq \mu_2$ and $\nu_1 \leq \nu_2$, it is clear that

$$\pi_1(E \times F) \leq \pi_2(E \times F)$$

for every measurable rectangle $E \times F$, hence for every finite disjoint union of measurable rectangles, that is, $\pi_1 \leq \pi_2$ on \mathcal{A} . Let

$$\mathcal{U} = \{M \in \mathcal{S} \times \mathcal{T} : \pi_1(M) \leq \pi_2(M)\}.$$

By the preceding remarks, $\mathcal{A} \subset \mathcal{U}$; our problem is to show that $\mathcal{S} \times \mathcal{T} \subset \mathcal{U}$. By the lemma on monotone classes (4.6.6) we need only observe that \mathcal{U} is

¹Cf. the author, *Measure and integration* [Macmillan, New York, 1965; reprinted Chelsea, Bronx, NY, 1970], p. 32, Lemma 2.

a monotone class. Indeed, if (M_n) is a sequence in \mathcal{U} with $M_n \uparrow M$ or $M_n \downarrow M$, then

$$\pi_1(M_n) \leq \pi_2(M_n) \quad \text{for all } n,$$

and passage to the limit yields $\pi_1(M) \leq \pi_2(M)$, with finiteness playing a role in the case of decreasing sequences (2.6.2, 2.6.3), whence the claim.

Define $\pi : \mathcal{S} \times \mathcal{T} \rightarrow [0, +\infty]$ by the formula

$$\pi(M) = \sup_n \pi_n(M) \quad (M \in \mathcal{S} \times \mathcal{T});$$

briefly, $\pi_n \uparrow \pi$ on $\mathcal{S} \times \mathcal{T}$. From $(*)$ we know that $\pi(E \times F) = \mu(E)\nu(F)$ for all $E \in \mathcal{S}$, $F \in \mathcal{T}$, in particular $\pi(P_n \times Q_n) = \mu(P_n)\nu(Q_n) < +\infty$ for all n ; to complete the proof of existence, we need only show that π is a measure on $\mathcal{S} \times \mathcal{T}$.

At any rate, it is clear that $\pi \geq 0$ and that $\pi(\emptyset) = 0$. If M and N are disjoint sets in $\mathcal{S} \times \mathcal{T}$, then

$$\begin{aligned} \pi(M \cup N) &= \sup_n \pi_n(M \cup N) = \sup_n [\pi_n(M) + \pi_n(N)] \\ &= \sup_n \pi_n(M) + \sup_n \pi_n(N) = \pi(M) + \pi(N) \end{aligned}$$

(cf. §1.15, Exercise 4, (ii)), thus π is finitely additive. Finally, if (M_k) is an increasing sequence in $\mathcal{S} \times \mathcal{T}$ with union M , then $\pi(M_k) \uparrow \pi(M)$ follows from the computation

$$\begin{aligned} \pi(M) &= \sup_n \pi_n(M) = \sup_n \left(\sup_k \pi_n(M_k) \right) \\ &= \sup_k \left(\sup_n \pi_n(M_k) \right) = \sup_k \pi(M_k), \end{aligned}$$

the key step being the interchange of order of the suprema (§1.15, Exercise 7), thus π is a measure (§2.6, Exercise 3).

Uniqueness: If π' is another measure on $\mathcal{S} \times \mathcal{T}$ such that $\pi'(E \times F) = \mu(E)\nu(F)$ for all measurable rectangles $E \times F$, then (by finite additivity) $\pi' = \pi$ on the algebra \mathcal{A} they generate; since $\pi'|_{\mathcal{A}} = \pi|_{\mathcal{A}}$ is σ -finite, it follows from the unique extension theorem (7.1.12) that $\pi' = \pi$ on the σ -algebra $\mathcal{S} \times \mathcal{T}$ generated by \mathcal{A} . \diamond

7.2.13. Definition. With notations as in the preceding corollary, π is called the **product** of the σ -finite measures μ and ν (in that order), written $\pi = \mu \times \nu$, and $(X \times Y, \mathcal{S} \times \mathcal{T}, \mu \times \nu)$ is called the **product measure space**.

It is possible to prove that if μ and ν are arbitrary (not necessarily σ -finite) measures on \mathcal{S} and \mathcal{T} , then there exists a measure π on $\mathcal{S} \times \mathcal{T}$ such that $\pi(E \times F) = \mu(E)\nu(F)$ for all measurable rectangles $E \times F$. The

price for dropping σ -finiteness is high: (1) π is no longer unique (§7.1, Exercise 2), (2) the Fubini–Tonelli theorems of the next two sections, the capital application of product measure, are not in general valid for π , (3) the construction of π no longer rests on the utterly transparent finite case (7.2.11), and (4) to construct π one is forced to contemplate ‘measurable’ functions with values $+\infty$ on sets of positive measure.² Absent a compelling application for the general case, the price is too high. (The case that one of μ and ν is σ -finite is worked out in the exercises.)

7.2.14. *Remark.* Suppose (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) are both equal to the Lebesgue measure space $(\mathbb{R}, \mathcal{M}, \lambda)$ of 2.4.13, let S be a non-measurable subset of \mathbb{R} (2.5.1) and let $T = S \times \{1\} \subset \mathbb{R} \times \mathbb{R}$. Then $T \subset \mathbb{R} \times \{1\}$, where $(\lambda \times \lambda)(\mathbb{R} \times \{1\}) = 0$, but $T \notin \mathcal{M} \times \mathcal{M}$ (because the section $T^1 = S$ is not measurable), thus the measure space $(\mathbb{R}^2, \mathcal{M} \times \mathcal{M}, \lambda \times \lambda)$ is not complete in the sense of §2.4, Exercise 8.

By *planar Lebesgue measure* $\lambda^{(2)}$ on \mathbb{R}^2 one usually means the restriction of the outer measure $(\lambda \times \lambda)^*$ to the σ -algebra of $(\lambda \times \lambda)^*$ -measurable subsets of \mathbb{R}^2 (cf. Example 2.4.13, (ii)); for planar Lebesgue measure, T is measurable and $\lambda^{(2)}(T) = 0$. Thus, the (complete) measure $\lambda^{(2)}$ is a proper extension of the product measure $\lambda \times \lambda$.

Exercises

1. (i) If μ and ν are real (complex) measures on \mathcal{S} and \mathcal{T} , respectively, then there exists a unique real (complex) measure π on $\mathcal{S} \times \mathcal{T}$ such that $\pi(E \times F) = \mu(E)\nu(F)$ for all measurable rectangles. One writes $\pi = \mu \times \nu$.

(ii) If, for example, $\mathcal{M}_{\mathbb{C}}(\mathcal{S})$ denotes the set of all complex measures on \mathcal{S} , then $\mathcal{M}_{\mathbb{C}}(\mathcal{S})$ is a complex vector space for the ‘pointwise’ linear operations (6.5.2) and the correspondence $(\mu, \nu) \mapsto \mu \times \nu$ is a bilinear mapping $\mathcal{M}_{\mathbb{C}}(\mathcal{S}) \times \mathcal{M}_{\mathbb{C}}(\mathcal{T}) \rightarrow \mathcal{M}_{\mathbb{C}}(\mathcal{S} \times \mathcal{T})$:

$$\mu \times (\nu_1 + \nu_2) = \mu \times \nu_1 + \mu \times \nu_2,$$

$$(\mu_1 + \mu_2) \times \nu = \mu_1 \times \nu + \mu_2 \times \nu,$$

$$\mu \times (c\nu) = c(\mu \times \nu) = (c\mu) \times \nu.$$

{Hint: (i) Existence: 4.8.8, 6.5.2. Uniqueness: Revisit the proof of Lemma 7.1.8.}

2. Let (X, \mathcal{S}, μ) , (Y, \mathcal{T}, ν) , (Z, \mathcal{U}, ρ) be σ -finite measure spaces. The following assertions (iii)–(ix) depend heavily on the uniqueness that comes along with products of σ -finite measures.

²Cf. H.L. Royden, *Real analysis* [3rd edn., Macmillan, New York, 1988], p. 303ff; R.G. Bartle, *The elements of integration* [Wiley, New York, 1966], Chapter 10; H.S. Bear, *A primer of Lebesgue integration* [Academic Press, New York, 1995], Chapter 14.

(i) If $P \in \mathcal{S}$, define $\mu_P : \mathcal{S} \rightarrow [0, +\infty]$ by the formula $\mu_P(E) = \mu(P \cap E)$ (cf. 4.8.4). Then μ_P is a measure on \mathcal{S} , finite if and only if $\mu(P) < \infty$.

(ii) If (P_n) is a sequence in \mathcal{S} such that $P_n \uparrow X$, then (μ_{P_n}) is an increasing sequence of measures such that $\mu(E) = \sup_n \mu_{P_n}(E)$ for all $E \in \mathcal{S}$; expressed concisely, $\mu_{P_n} \uparrow \mu$.

(iii) If $P \in \mathcal{S}$ and $Q \in \mathcal{T}$ then $(\mu \times \nu)_{P \times Q} = \mu_P \times \nu_Q$.

(iv) If $(P_n), (Q_n)$ are sequences in \mathcal{S}, \mathcal{T} such that $P_n \uparrow X$ and $Q_n \uparrow Y$, then $\mu_{P_n} \times \nu_{Q_n} \uparrow \mu \times \nu$.

(v) The set \mathcal{A} of all finite disjoint unions of sets $(E \times F) \times G$ ($E \in \mathcal{S}$, $F \in \mathcal{T}$, $G \in \mathcal{U}$) is an algebra of subsets of $(X \times Y) \times Z$, and the σ -algebra generated by \mathcal{A} is $(\mathcal{S} \times \mathcal{T}) \times \mathcal{U}$.

(vi) $(\mu \times \nu) \times \rho$ is the unique measure π on $(\mathcal{S} \times \mathcal{T}) \times \mathcal{U}$ such that $\pi((E \times F) \times G) = \mu(E)\nu(F)\rho(G)$ for all $E \in \mathcal{S}, F \in \mathcal{T}, G \in \mathcal{U}$.

(vii) There exists a unique measure π on the σ -algebra $\mathcal{S} \times \mathcal{T} \times \mathcal{U}$ of subsets of $X \times Y \times Z$ generated by the sets $E \times F \times G$ ($E \in \mathcal{S}, F \in \mathcal{T}, G \in \mathcal{U}$) such that

$$\pi(E \times F \times G) = \mu(E)\nu(F)\rho(G)$$

for all E, F, G . One writes $\pi = \mu \times \nu \times \rho$.

(viii) The natural bijection $(X \times Y) \times Z \rightarrow X \times (Y \times Z)$ transforms $(\mathcal{S} \times \mathcal{T}) \times \mathcal{U}$ into $\mathcal{S} \times (\mathcal{T} \times \mathcal{U})$, and the measure $(\mu \times \nu) \times \rho$ into $\mu \times (\nu \times \rho)$.

(ix) With the natural identifications of $(X \times Y) \times Z$, $X \times Y \times Z$ and $X \times (Y \times Z)$, one has the 'associative law'

$$(\mu \times \nu) \times \rho = \mu \times \nu \times \rho = \mu \times (\nu \times \rho).$$

{Hints: (i), (ii) Here μ need not be σ -finite. (iii) Evaluate the two measures at a measurable rectangle $E \times F$. (iv) Immediate from (iii) and (ii).

(v) If $G \in \mathcal{U}$ and $\mathcal{S}(\mathcal{A})$ is the σ -algebra generated by \mathcal{A} , then the set $\{M \in \mathcal{S} \times \mathcal{T} : M \times G \in \mathcal{S}(\mathcal{A})\}$ is a monotone class containing the algebra generated by the sets $E \times F$ ($E \in \mathcal{S}, F \in \mathcal{T}$).

(vi) Let π be a measure with the indicated property. If $\mu(P) < +\infty$, $\nu(Q) < +\infty$ and $\rho(R) < +\infty$, observe that $\pi_{(P \times Q) \times R} = ((\mu \times \nu) \times \rho)_{(P \times Q) \times R}$ on the algebra \mathcal{A} of (v), hence on $(\mathcal{S} \times \mathcal{T}) \times \mathcal{U}$. Complete the proof by applying (iv).

(vii) The natural bijection $(X \times Y) \times Z \rightarrow X \times Y \times Z$ transports $(\mathcal{S} \times \mathcal{T}) \times \mathcal{U}$ into $\mathcal{S} \times \mathcal{T} \times \mathcal{U}$ and $(\mu \times \nu) \times \rho$ into a measure π with the indicated property.}

3. (i) If (μ_n) is a sequence of measures on a σ -algebra \mathcal{S} such that $\mu_1 \leq \mu_2 \leq \mu_3 \leq \dots$, then the formula

$$\mu(E) = \sup_n \mu_n(E) \quad (E \in \mathcal{S})$$

defines a measure μ on \mathcal{S} . One writes $\mu_n \uparrow \mu$.

(ii) Generalize (i) to increasing nets $(\mu_i)_{i \in I}$ of measures on \mathcal{S} .

(iii) Discuss the sum $\mu = \sum_{i \in I} \mu_i$ of an arbitrary family $(\mu_i)_{i \in I}$ of measures on \mathcal{S} , by contemplating the net of finite subsums of the family.

{Hint: (i) Associativity of sups (see the proof of 7.2.12). (ii), (iii) Cf. §1.15, Exercise 4.³}

4. If μ_1 and μ_2 are σ -finite measures on \mathcal{S} such that $\mu_1 \leq \mu_2$, and if ν_1, ν_2 are σ -finite measures on \mathcal{T} such that $\nu_1 \leq \nu_2$, then $\mu_1 \times \nu_1 \leq \mu_2 \times \nu_2$ on $\mathcal{S} \times \mathcal{T}$.

{Hint: §7.1, Exercise 3.}

5. (i) Suppose ν is a *finite* measure on \mathcal{T} and μ is an arbitrary measure on \mathcal{S} . Then, for every $M \in \mathcal{S} \times \mathcal{T}$, $0 \leq f_M \leq \nu(Y)$ and f_M is measurable with respect to \mathcal{S} ; writing $\int f_M d\mu$ as authorized by 4.5.6 (with value $+\infty$ if f_M is not integrable), the formula

$$\pi(M) = \int f_M d\mu$$

defines a measure π on $\mathcal{S} \times \mathcal{T}$ such that $\pi(E \times F) = \mu(E)\nu(F)$ for every measurable rectangle $E \times F$. Trespassing a little on the turf of Definition 7.2.13, let us write $\pi = \mu \times \nu$.

(ii) Let ν_1 and ν_2 be *finite* measures on \mathcal{T} , let μ be an arbitrary measure on \mathcal{S} , and construct $\mu \times \nu_1$ and $\mu \times \nu_2$ as in (i). If $\nu_1 \leq \nu_2$ then $\mu \times \nu_1 \leq \mu \times \nu_2$.

(iii) If ν is a σ -finite measure on \mathcal{T} and if μ is an arbitrary measure on \mathcal{S} , then there exists a measure π on $\mathcal{S} \times \mathcal{T}$ such that $\pi(E \times F) = \mu(E)\nu(F)$ for all measurable rectangles $E \times F$.

(iv) If μ_1, μ_2 are measures on \mathcal{S} such that $\mu_1 \leq \mu_2$ and if ν_1, ν_2 are *finite* measures on \mathcal{T} such that $\nu_1 \leq \nu_2$, then $\mu_1 \times \nu_1 \leq \mu_2 \times \nu_2$, where the 'product measures' are constructed as in (i).

{Hints: (i) $f_{E \times F} = \nu(F)\varphi_E$, and §4.5, Exercise 1.

(ii) If $M \in \mathcal{S} \times \mathcal{T}$, then $\nu_1(M_x) \leq \nu_2(M_x)$ for all $x \in X$.

(iii) Let (ν_n) be a sequence of finite measures on \mathcal{T} such that $\nu_n \uparrow \nu$, form the measures $\mu \times \nu_n$ as in (i), and contemplate (ii) and Exercise 3, (i).}

6. Let (X, \mathcal{S}, μ) be any measure space, $f : X \rightarrow \mathbb{R}$ a measurable function such that $f \geq 0$, and let

$$M = \{(x, y) : x \in X, 0 \leq y < f(x)\}$$

(so to speak, M is the area under the graph of f). Let \mathcal{B} be the σ -algebra of Borel sets in \mathbb{R} , λ the restriction of Lebesgue measure to \mathcal{B} , and let π be a measure on $\mathcal{S} \times \mathcal{B}$ such that $\pi(E \times B) = \mu(E)\lambda(B)$ for all $E \in \mathcal{S}$, $B \in \mathcal{B}$ (Exercise 5, (iii)).

³Cf. the author, *op. cit.*, p. 31, §10.

Show that $M \in \mathcal{S} \times \mathcal{B}$ and $\pi(M) = \int f d\mu$ (possibly $+\infty$; cf. 4.5.6). Thus $f \in \mathcal{L}^1(\mu) \Leftrightarrow \pi(M) < +\infty$.

{Hint: Let (f_n) be a sequence of simple functions such that $0 \leq f_n \uparrow f$ and let $M_n = \{(x, y) : x \in X, 0 \leq y < f_n(x)\}$. Note that $M_n \uparrow M$, each M_n is a finite union of measurable rectangles, and $\pi(M_n)$ is a bounded sequence if and only if the f_n are integrable simple functions with bounded integrals.}

7.3. Iterated Integrals, Fubini–Tonelli Theorem for Finite Measures

The central theme of this section is the integration of functions with respect to the product measure $\pi = \mu \times \nu$ in the context of finite measure spaces (X, \mathcal{S}, μ) , (Y, \mathcal{T}, ν) (in the next section we advance to σ -finite measures). To motivate the new ideas that are needed, consider the case of the characteristic function φ_M of a set $M \in \mathcal{S} \times \mathcal{T}$. Citing the definition of π in the preceding section (proof of Theorem 7.2.11), we have

$$\begin{aligned} \int \varphi_M d\pi &= \pi(M) = \int f_M d\mu \\ &= \int \nu(M_x) d\mu(x) \\ &= \int \left(\int \varphi_{M_x} d\nu \right) d\mu(x) \\ &= \int \left(\int \varphi_{M_x}(y) d\nu(y) \right) d\mu(x); \end{aligned}$$

now,

$$\varphi_{M_x}(y) = \begin{cases} 1 & \text{if } y \in M_x \\ 0 & \text{if } y \notin M_x \end{cases} = \begin{cases} 1 & \text{if } (x, y) \in M \\ 0 & \text{if } (x, y) \notin M \end{cases} = \varphi_M(x, y),$$

thus

$$(1) \quad \int \varphi_M d\pi = \int \left(\int \varphi_M(x, y) d\nu(y) \right) d\mu(x).$$

A similar computation, based on the formula $\pi(M) = \int g^M d\nu$ (cf. 7.2.11) yields

$$(2) \quad \int \varphi_M d\pi = \int \left(\int \varphi_M(x, y) d\mu(x) \right) d\nu(y).$$

So to speak, the ‘double integral’

$$\int \varphi_M d\pi = \int \varphi_M(x, y) d\pi(x, y)$$

is equal to the 'iterated integrals' appearing on the right sides of (1) and (2); concisely,

$$(*) \quad \int \varphi_M d\pi = \iint \varphi_M d\nu d\mu = \iint \varphi_M d\mu d\nu.$$

Our goal in this section is to prove $(*)$ with φ_M replaced by an arbitrary π -integrable function h , but first we must give an appropriate definition of the symbols $\iint h d\nu d\mu$ and $\iint h d\mu d\nu$; the first step is to discuss the 'partial' functions $y \mapsto h(x, y)$ and $x \mapsto h(x, y)$ obtained by fixing one variable and letting the other vary:

7.3.1. Definition. Let $h : X \times Y \rightarrow Z$, where X, Y, Z are arbitrary nonempty sets. For each $x \in X$, the x -**section** of h is the function $h_x : Y \rightarrow Z$ defined by the formula

$$h_x(y) = h(x, y) \quad (y \in Y),$$

and, for each $y \in Y$, the y -**section** of h is the function $h^y : X \rightarrow Z$ defined by

$$h^y(x) = h(x, y) \quad (x \in X).$$

7.3.2. Example. (i) Let $i : X \times Y \rightarrow X \times Y$ be the identity mapping on $X \times Y$. If $x \in X$ then

$$i_x(y) = i(x, y) = (x, y) \quad \text{for all } y \in Y,$$

and if $y \in Y$ then

$$i^y(x) = i(x, y) = (x, y) \quad \text{for all } x \in X,$$

thus the sections i_x and i^y of the identity mapping on $X \times Y$ are precisely the functions introduced in Definition 7.2.4. Note, incidentally, that if $h : X \times Y \rightarrow Z$ as in 7.3.1, then $h_x = h \circ i_x$ and $h^y = h \circ i^y$ for all $x \in X$ and $y \in Y$; for example,

$$h_x(y) = h(x, y) = h(i_x(y)) = (h \circ i_x)(y)$$

for all $y \in Y$.

(ii) If $M \subset X \times Y$ and $\varphi_M : X \times Y \rightarrow \{0, 1\}$ is the characteristic function of M then, by the computation at the beginning of the section,

$$\varphi_{M_x}(y) = \varphi_M(x, y) = (\varphi_M)_x(y) \quad \text{for all } x \in X \text{ and } y \in Y,$$

thus $(\varphi_M)_x = \varphi_{M_x}$ for all $x \in X$, and similarly $(\varphi_M)^y = \varphi_{M^y}$ for all $y \in Y$.

(iii) If (X, \mathcal{S}) and (Y, \mathcal{T}) are measurable spaces and $h : X \times Y \rightarrow \mathbb{R}$ (or \mathbb{C}) is measurable with respect to $\mathcal{S} \times \mathcal{T}$, then every section of h is measurable.

{Proof: Suppose h is real-valued. For example, given any $x \in X$ and any Borel set B in \mathbb{R} , we are to show that $h_x^{-1}(B) \in \mathcal{T}$ (4.1.3). At any

rate, we know from the measurability of h that $h^{-1}(B) \in \mathcal{S} \times \mathcal{T}$. As noted in (i) above, $h_x = h \circ i_x$, therefore the set

$$h_x^{-1}(B) = i_x^{-1}(h^{-1}(B)) = (h^{-1}(B))_x$$

belongs to \mathcal{T} by Lemma 7.2.6. The case that h is complex-valued then follows easily from the observation that $\operatorname{Re} h_x = (\operatorname{Re} h)_x$ and $\operatorname{Im} h_x = (\operatorname{Im} h)_x$ (and similarly for y -sections).}

(iv) If (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) are finite measure spaces, $\pi = \mu \times \nu$ and $h \in \mathcal{L}^1(\pi)$, it does not follow that every section of h is integrable (Exercise 2).

Though our ultimate concern is with finite and σ -finite measure spaces, the following concept makes sense for arbitrary measure spaces:

7.3.3. Definition. Let (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) be measure spaces and let $h : X \times Y \rightarrow \mathbb{C}$. We say that the **iterated integral**

$$\iint h d\nu d\mu$$

exists if there exist a set $E \in \mathcal{S}$ with $\mu(E) = 0$ and a function $f \in \mathcal{L}^1_{\mathbb{C}}(\mu)$ such that

$$x \in X - E \quad \Rightarrow \quad h_x \in \mathcal{L}^1_{\mathbb{C}}(\nu) \quad \text{and} \quad \int h_x d\nu = f(x);$$

we then write

$$\iint h d\nu d\mu = \int f d\mu$$

(clearly well-defined, that is, independent of the particular pair E, f with the indicated properties). Informally: almost all of the x -sections are integrable and their integrals determine a function that is integrable with respect to μ ; suggestively,

$$\iint h d\nu d\mu = \int \left(\int h(x, y) d\nu(y) \right) d\mu(x).$$

Similarly, we say that the iterated integral

$$\iint h d\mu d\nu$$

exists if there exist a set $F \in \mathcal{T}$ and a function $g \in \mathcal{L}^1_{\mathbb{C}}(\nu)$ such that

$$y \in Y - F \quad \Rightarrow \quad h^y \in \mathcal{L}^1_{\mathbb{C}}(\mu) \quad \text{and} \quad \int h^y d\mu = g(y);$$

and we then write $\iint h d\mu d\nu = \int g d\nu$.

7.3.4. Remarks. (i) If μ and ν are finite then, for every $M \in \mathcal{S} \times \mathcal{T}$, the iterated integrals of φ_M both exist and are equal to $\int \varphi_M d(\mu \times \nu) =$

$(\mu \times \nu)(M)$ by the remarks at the beginning of the section (and both of the sets E, F in the foregoing definition can be taken to be the empty set).

(ii) The foregoing definition is 'linear' in the following sense: If h and k are functions such that $\iint h d\nu d\mu$ and $\iint h d\mu d\nu$ both exist, and if a, b are complex numbers, then $\iint (ah + bk) d\nu d\mu$ exists and

$$\iint (ah + bk) d\nu d\mu = a \iint h d\nu d\mu + b \iint k d\nu d\mu.$$

{Reason: The union of two μ -null sets is μ -null, a linear combination of μ -integrable functions is μ -integrable, and $f \mapsto \int f d\mu$ is a linear form on $\mathcal{L}_\mathbb{C}^1(\mu)$.}

(iii) In the foregoing definition, it is not assumed that h is measurable with respect to $\mathcal{S} \times \mathcal{T}$.

For the rest of the section we restrict attention to *finite* measures.

7.3.5. Lemma. Assume (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) are finite measure spaces. If $h : X \times Y \rightarrow \mathbb{C}$ is a simple function with respect to $\mathcal{S} \times \mathcal{T}$ (hence is integrable with respect to $\mu \times \nu$), then every section of h is integrable, both iterated integrals of h exist, and

$$\iint h d\nu d\mu = \iint h d\mu d\nu = \int h d(\mu \times \nu).$$

Proof. By linearity (Remark 7.3.4, (ii)) we can suppose that h is the characteristic function of a set $M \in \mathcal{S} \times \mathcal{T}$, a case that is covered by Remark 7.3.4, (i). Explicitly, if $h = \sum_{k=1}^n c_k \varphi_{M_k}$ then, for all $x \in X$ and $y \in Y$,

$$h_x = \sum_{k=1}^n c_k \varphi_{(M_k)_x}, \quad h^y = \sum_{k=1}^n c_k \varphi_{(M_k)^y},$$

and

$$\int h_x d\nu = \sum_{k=1}^n c_k f_{M_k}(x), \quad \int h^y d\mu = \sum_{k=1}^n c_k g^{M_k}(y),$$

thus the requirements of Definition 7.3.3 are met with $E = \emptyset$, $f = \sum_{k=1}^n c_k f_{M_k}$, $F = \emptyset$, $g = \sum_{k=1}^n c_k g^{M_k}$. \diamond

The following theorem will be generalized to σ -finite measures in the next section:

7.3.6. Theorem. (Fubini-Tonelli theorem,¹ finite case) Let (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) be finite measure spaces, $h : X \times Y \rightarrow \mathbb{R}$. If $h \geq 0$ and h is measurable with respect to $\mathcal{S} \times \mathcal{T}$, then the following conditions are equivalent:

(a) h is integrable with respect to $\mu \times \nu$;

¹Guido Fubini (1879–1943), Leonida Tonelli (1885–1946).

(b) $\iint h d\nu d\mu$ exists;

(c) $\iint h d\mu d\nu$ exists.

When this is the case, $\int h d(\mu \times \nu) = \iint h d\nu d\mu = \iint h d\mu d\nu$.

Proof. Let (h_n) be a sequence of simple functions (with respect to $\mathcal{S} \times \mathcal{T}$) such that $0 \leq h_n \uparrow h$. Every section of h_n is simple (with respect to \mathcal{S} or \mathcal{T} , as the case may be), hence integrable (with respect to μ or ν), and $0 \leq (h_n)_x \uparrow h_x$, $0 \leq (h_n)^y \uparrow h^y$ for all $x \in X$, $y \in Y$. Define

$$f_n(x) = \int (h_n)_x d\nu, \quad g_n(y) = \int (h_n)^y d\mu$$

for $x \in X$, $y \in Y$. As noted in the proof of the lemma, $f_n \in \mathcal{L}^1(\mu)$, $g_n \in \mathcal{L}^1(\nu)$ and

$$(*) \quad \int f_n d\mu = \int g_n d\nu = \int h_n d(\mu \times \nu) \quad \text{for all } n.$$

Since $(h_n)_x \uparrow$ for each x , and $(h_n)^y \uparrow$ for each y , we have

$$f_n \uparrow \quad \text{and} \quad g_n \uparrow$$

by the monotonicity of integration.

(a) \Rightarrow (b): Suppose $h \in \mathcal{L}^1(\mu \times \nu)$. Then

$$\int h_n d(\mu \times \nu) \uparrow \int h d(\mu \times \nu)$$

(by Definition 4.4.3, or by the monotone convergence theorem, 4.5.3), thus

$$\int f_n d\mu \uparrow \int h d(\mu \times \nu) < +\infty$$

by (*). Since (f_n) is an increasing sequence, it follows from the monotone convergence theorem (for μ) that there exists $f \in \mathcal{L}^1(\mu)$ such that $f_n \uparrow f$ a.e. (for μ) and

$$\int f d\mu = \sup_n \int f_n d\mu = \int h d(\mu \times \nu).$$

Choose $E \in \mathcal{S}$, $\mu(E) = 0$ so that $f_n \uparrow f$ on $X - E$; then, for all $x \in X - E$,

$$\sup_n \int (h_n)_x d\nu = \sup_n f_n(x) = f(x) < +\infty,$$

and since $(h_n)_x \uparrow h_x$ and h_x is measurable (with respect to \mathcal{T}), it follows from the monotone convergence theorem (for ν) that $h_x \in \mathcal{L}^1(\nu)$ and

$$\int h_x d\nu = \sup_n \int (h_n)_x d\nu = f(x).$$

The pair E, f meets the requirements of Definition 7.3.3, thus the iterated integral $\iint h d\nu d\mu$ exists and

$$\iint h d\nu d\mu = \int f d\mu = \int h d(\mu \times \nu).$$

(a) \Rightarrow (c): Similarly, with $\iint h d\mu d\nu = \int h d(\mu \times \nu)$.

(b) \Rightarrow (a): Suppose $\iint h d\nu d\mu$ exists, and let E, f be as in Definition 7.3.3. For each $x \in X - E$ we have $0 \leq (h_n)_x \uparrow h_x$ and

$$\sup_n \int (h_n)_x d\nu = \int h_x d\nu < +\infty$$

by the monotone convergence theorem, in other words

$$f_n(x) \uparrow f(x).$$

Thus $f_n \uparrow f$ a.e. (for μ) and $\int f_n d\mu \uparrow \int f d\mu$ by the monotone convergence theorem; in view of (*), we see that the sequence

$$\int h_n d(\mu \times \nu) = \int f_n d\mu$$

is bounded, whence $h \in \mathcal{L}^1(\mu \times \nu)$ by the definition of integrability (4.4.3).

(c) \Rightarrow (a): Similarly. \diamond

7.3.7. Corollary. (Fubini's theorem for finite measures) *Let (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) be finite measure spaces. If $h \in \mathcal{L}^1_{\mathbb{C}}(\mu \times \nu)$ then both iterated integrals of h exist and are equal to $\int h d(\mu \times \nu)$.*

Proof. By linearity (7.3.4, (ii)) we can suppose that $h \geq 0$, then cite the theorem. \diamond

The following corollary will be useful in the next section for generalizing to the σ -finite case:

7.3.8. Corollary. *Let (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) be finite measure spaces. If $h : X \times Y \rightarrow \mathbb{C}$ is bounded and measurable (with respect to $\mathcal{S} \times \mathcal{T}$), then every section of h is integrable and, defining*

$$f(x) = \int h_x d\nu, \quad g(y) = \int h^y d\mu$$

for all $x \in X$ and $y \in Y$, we have $f \in \mathcal{L}^1(\mu)$, $g \in \mathcal{L}^1(\nu)$ and

$$\iint h d\nu d\mu = \iint h d\mu d\nu = \int f d\mu = \int g d\nu = \int h d(\mu \times \nu).$$

Proof. We can suppose by linearity that $h \geq 0$ (7.3.4, (ii)). At any rate, we know that h is integrable with respect to the finite measure $\mu \times \nu$ (4.4.5). For every $x \in X$, h_x is bounded and measurable with respect

to \mathcal{T} (7.3.2, (iii)), therefore integrable with respect to ν ; similarly for y -sections. Thus f and g are defined everywhere on X and Y , respectively.

Note that f and g are bounded. For, if $0 \leq h \leq K < +\infty$, then $\int h_x d\nu \leq K\nu(Y) < +\infty$ for all $x \in X$, thus

$$0 \leq f \leq K\nu(Y) < +\infty$$

and similarly $0 \leq g \leq K\mu(X) < +\infty$.

The case that h is simple is covered by Lemma 7.3.5. In general, let (h_n) be a sequence of simple functions (with respect to $\mathcal{S} \times \mathcal{T}$) such that $0 \leq h_n \uparrow h$. Define

$$f_n(x) = \int (h_n)_x d\nu \quad (x \in X);$$

we know that $f_n \in \mathcal{L}^1(\mu)$, in particular f_n is measurable with respect to \mathcal{S} . For each $x \in X$,

$$(h_n)_x \uparrow h_x \in \mathcal{L}^1(\nu),$$

and integration with respect to ν yields $f_n(x) \uparrow f(x)$. Thus $f_n \uparrow f$ pointwise. Since the f_n are measurable, so is their pointwise (real-valued) limit f ; also, f is bounded, so $f \in \mathcal{L}^1(\mu)$. Similarly, $g \in \mathcal{L}^1(\nu)$, and the asserted formulas are immediate from the theorem. \diamond

Exercises

1. If $f : X \rightarrow Y$, $\varphi_B : Y \rightarrow \{0, 1\}$ is the characteristic function of a set $B \subset Y$, and $\varphi_B \circ f : X \rightarrow \{0, 1\}$ is the composite function, then $\varphi_B \circ f = \varphi_{f^{-1}(B)}$.

2. Let λ_0 be Lebesgue measure on the closed interval $[0, 1]$ (Example 4.7.6), defined on the σ -algebra \mathcal{M}_0 of Lebesgue-measurable sets contained in $[0, 1]$, and let $\pi = \lambda_0 \times \lambda_0$. Let $h : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$ be the function such that $h(x, 0) = 1/x$ for $0 < x \leq 1$ and $h(x, y) = 0$ for all other (x, y) .

(i) h is measurable with respect to $\mathcal{M}_0 \times \mathcal{M}_0$ (indeed, with respect to $\mathcal{B}_0 \times \mathcal{B}_0$, where \mathcal{B}_0 is the σ -algebra of Borel sets contained in $[0, 1]$).

(ii) h is π -integrable.

(iii) The section h^0 of h is not λ_0 -integrable.

{Hint: (i) Apply criterion (b) of 4.1.8; for example, if $c > 0$ then $\{(x, y) : h(x, y) > c\} = (0, 1/c) \times (0, 1]$. (ii) $h^0(x) = 1/x$ for $x > 0$.}

7.4. Fubini–Tonelli Theorem for σ -Finite Measures

Throughout this section, (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) are σ -finite measure spaces, that is, μ and ν are σ -finite measures on the σ -algebras \mathcal{S} and \mathcal{T}

(7.1.9). Let $(X \times Y, \mathcal{S} \times \mathcal{T}, \mu \times \nu)$ be the product measure space (7.2.13); as observed in the course of its construction, the measure $\mu \times \nu$ is also σ -finite (7.2.12).

Our objective is to extend the results of the preceding section (for finite measures) to the σ -finite case. The strategy is the same as in the construction of $\mu \times \nu$: express each of μ and ν as the supremum of a suitable increasing sequence of finite measures, let the finite case do all of the hard work, then ‘pass to the limit’ via some easy technical lemmas.

As in the proof of 7.2.12, let (P_n) be a sequence of sets in \mathcal{S} such that $P_n \uparrow X$ and $\mu(P_n) < +\infty$ for all n , and let (Q_n) be a sequence of sets in \mathcal{T} such that $Q_n \uparrow Y$ and $\nu(Q_n) < +\infty$ for all n . Then

$$(\mu \times \nu)(P_n \times Q_n) = \mu(P_n)\nu(Q_n) < +\infty$$

for all n . The formulas

$$\mu_n(E) = \mu(P_n \cap E) \quad (E \in \mathcal{S})$$

$$\nu_n(F) = \nu(Q_n \cap F) \quad (F \in \mathcal{T})$$

define finite measures μ_n on \mathcal{S} and ν_n on \mathcal{T} such that $\mu_n \uparrow \mu$ and $\nu_n \uparrow \nu$. The measures $\mu_n \times \nu_n$ are finite (7.2.11).

7.4.1. Lemma. $\mu_n \times \nu_n \uparrow \mu \times \nu$.

Proof. In view of the definition of $\mu \times \nu$ (7.2.13), the argument for this is given in the proof of Corollary 7.2.12. \diamond

The following notation, already employed in 4.8.4 in connection with finite signed measures, will also be useful in the present context:

7.4.2. *Definition.* For each measurable set $A \in \mathcal{S}$, define a set function $\mu_A : \mathcal{S} \rightarrow [0, +\infty]$ by the formula

$$\mu_A(E) = \mu(A \cap E) \quad (E \in \mathcal{S}).$$

(One calls μ_A the **contraction** of μ to A .)

7.4.3. *Remarks.* (i) The set function μ_A introduced in the foregoing definition is a measure, by a straightforward argument that is evidently valid when μ is replaced by an arbitrary (not necessarily σ -finite) measure.

(ii) Every contraction of the σ -finite measure μ is also σ -finite. In order that μ_A be a finite measure, it is necessary and sufficient that $\mu(A) < +\infty$.

(iii) With the notations introduced at the beginning of the section, $\mu_n = \mu_{P_n}$ and $\nu_n = \nu_{Q_n}$.

7.4.4. Lemma. Suppose $A \in \mathcal{S}$ and $f : X \rightarrow \mathbb{C}$ is measurable with respect to \mathcal{S} (6.4.1). Then

$$f \in \mathcal{L}^1_{\mathbb{C}}(\mu_A) \quad \Leftrightarrow \quad f\varphi_A \in \mathcal{L}^1_{\mathbb{C}}(\mu),$$

and in this case

$$\int f d\mu_A = \int f\varphi_A d\mu.$$

Proof (valid for any measure space). By linearity, we can suppose without loss of generality that $f \geq 0$. {The crux of the matter is that $\operatorname{Re}(f\varphi_A) = (\operatorname{Re} f)\varphi_A$ and $\operatorname{Im}(f\varphi_A) = (\operatorname{Im} f)\varphi_A$; and, when f is real-valued, that $(f\varphi_A)_+ = f_+\varphi_A$ and $(f\varphi_A)_- = f_-\varphi_A$.}

case 1: $f = \varphi_E$ for some $E \in \mathcal{S}$.

Then

$$\begin{aligned} f \in \mathcal{L}^1(\mu_A) &\Leftrightarrow \mu_A(E) < +\infty \\ &\Leftrightarrow \mu(A \cap E) < +\infty \\ &\Leftrightarrow \varphi_{A \cap E} \in \mathcal{L}^1(\mu) \\ &\Leftrightarrow \varphi_A \varphi_E \in \mathcal{L}^1(\mu) \\ &\Leftrightarrow f\varphi_A \in \mathcal{L}^1(\mu) \end{aligned}$$

and in this case we have

$$\int f d\mu_A = \mu_A(E) = \mu(A \cap E) = \int \varphi_{A \cap E} d\mu = \int f\varphi_A d\mu.$$

case 2: f simple (and ≥ 0).

Immediate from case 1 and linearity.

case 3: the general case ($f \geq 0$).

Let (f_n) be a sequence of simple functions such that $0 \leq f_n \uparrow f$ pointwise on X (4.1.26). Then

$$0 \leq f_n \varphi_A \uparrow f\varphi_A,$$

where the $f_n \varphi_A$ are simple functions. Citing the definition of integral (4.4.1, 4.4.3) and case 2, we have

$$\begin{aligned} f \in \mathcal{L}^1(\mu_A) &\Leftrightarrow f_n \in \mathcal{L}^1(\mu_A) \quad (\forall n) \quad \text{and} \quad \int f_n d\mu_A \text{ is bounded} \\ &\Leftrightarrow f_n \varphi_A \in \mathcal{L}^1(\mu) \quad (\forall n) \quad \text{and} \quad \int f_n \varphi_A d\mu \text{ is bounded} \\ &\Leftrightarrow f\varphi_A \in \mathcal{L}^1(\mu), \end{aligned}$$

and in this case,

$$\int f d\mu_A = \sup_n \int f_n d\mu_A = \sup_n \int f_n \varphi_A d\mu = \int f\varphi_A d\mu. \quad \diamond$$

7.4.5. Lemma. Let (f_n) be a sequence of functions on X such that, for every n , $0 \leq f_n \in \mathcal{L}^1(\mu_n)$ and $f_n \leq f_{n+1}$ μ -a.e. The following

conditions are equivalent:

(a) the sequence $\int f_n d\mu_n$ is bounded;

(b) there exists a function $f \in \mathcal{L}^1(\mu)$ such that $f_n \uparrow f$ μ -a.e.

When this is the case,

$$\int f d\mu = \sup_n \int f_n d\mu_n.$$

Proof. Note that $f_n(x)$ is an increasing sequence for μ -almost every $x \in X$, briefly $f_n \uparrow \mu$ -a.e. Since $\mu_n = \mu_{P_n}$, by the preceding lemma we have $f_n \varphi_{P_n} \in \mathcal{L}^1(\mu)$ and

$$\int f_n \varphi_{P_n} d\mu = \int f_n d\mu_{P_n} = \int f_n d\mu_n;$$

also $f_n \varphi_{P_n} \uparrow \mu$ -a.e. By the monotone convergence theorem for μ , in order that there exist a function $f \in \mathcal{L}^1(\mu)$ such that $f_n \varphi_{P_n} \uparrow f$ μ -a.e. (equivalently $f_n \uparrow f$ μ -a.e., since, for each $x \in X$, $\varphi_{P_n}(x) = 1$ ultimately), it is necessary and sufficient that

$$\sup_n \int f_n \varphi_{P_n} d\mu < +\infty,$$

in other words

$$\sup_n \int f_n d\mu_n < +\infty;$$

when this is the case, we have

$$\sup_n \int f_n d\mu_n = \sup_n \int f_n \varphi_{P_n} d\mu = \int f d\mu. \quad \diamond$$

The analogue of the preceding lemma of course holds for the sequence $\nu_n \uparrow \nu$.

7.4.6. Lemma. If $A \in \mathcal{S}$ and $B \in \mathcal{T}$ then $(\mu \times \nu)_{A \times B} = \mu_A \times \nu_B$.

Proof. All measures in sight are σ -finite and, for every measurable rectangle $E \times F \in \mathcal{S} \times \mathcal{T}$,

$$\begin{aligned} (\mu \times \nu)_{A \times B}(E \times F) &= (\mu \times \nu)((A \times B) \cap (E \times F)) \\ &= (\mu \times \nu)((A \cap E) \times (B \cap F)) \\ &= \mu(A \cap E) \nu(B \cap F) \\ &= \mu_A(E) \nu_B(F) \\ &= (\mu_A \times \nu_B)(E \times F), \end{aligned}$$

therefore $(\mu \times \nu)_{A \times B} = \mu_A \times \nu_B$ by the uniqueness part of Corollary 7.2.12. \diamond

In particular, $(\mu \times \nu)_{P_n \times Q_n} = \mu_{P_n} \times \nu_{Q_n} = \mu_n \times \nu_n$, therefore the analogue of Lemma 7.4.5 also holds for the sequence $\mu_n \times \nu_n \uparrow \mu \times \nu$.

7.4.7. Theorem. (Fubini–Tonelli) If (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) are σ -finite measure spaces, and if h is a measurable function with respect to $\mathcal{S} \times \mathcal{T}$ such that $h \geq 0$ on $X \times Y$, then the following conditions are equivalent:

- (a) $h \in \mathcal{L}^1(\mu \times \nu)$;
- (b) $\iint d\nu d\mu$ exists;
- (c) $\iint h d\mu d\nu$ exists.

When this is the case,

$$\int h d(\mu \times \nu) = \iint h d\nu d\mu = \iint h d\mu d\nu.$$

Proof. The foregoing notations for the measures μ_n and ν_n are in force. Let (h_i) be a sequence of simple functions on $X \times Y$ (with respect to $\mathcal{S} \times \mathcal{T}$) such that $0 \leq h_i \uparrow h$. The h_i are of course integrable with respect to the finite measures $\mu_n \times \nu_n$; replacing h_i by $h_i \varphi_{P_i \times Q_i}$, we can suppose further that h_i is integrable with respect to $\mu \times \nu$. For each i and n , define

$$f_i^{(n)}(x) = \int (h_i)_x d\nu_n \quad (x \in X);$$

by Lemma 7.3.5 (or Corollary 7.3.8) we know that $f_i^{(n)} \in \mathcal{L}^1(\mu_n)$ and

$$(1) \quad \int f_i^{(n)} d\mu_n = \int h_i d(\mu_n \times \nu_n).$$

Note also that

$$(2) \quad \text{for each } n, \quad f_i^{(n)} \uparrow \text{ as } i \uparrow$$

(because, for each $x \in X$, $(h_i)_x \uparrow$ pointwise on Y), and that

$$(3) \quad \text{for each } i, \quad f_i^{(n)} \uparrow \text{ as } n \uparrow$$

(because (ν_n) is an increasing sequence of measures).

(a) \Rightarrow (b): Suppose $h \in \mathcal{L}^1(\mu \times \nu)$. Then, citing (1), we have

$$(*) \quad \int f_i^{(n)} d\mu_n = \int h_i d(\mu_n \times \nu_n) \leq \int h_i d(\mu \times \nu) \leq \int h d(\mu \times \nu) < +\infty$$

for all i and n .

Fix an index i . By (3), (*) and Lemma 7.4.5, there exists $f_i \in \mathcal{L}^1(\mu)$ such that $f_i^{(n)} \uparrow f_i$ μ -a.e., and

$$(4) \quad \begin{aligned} \int f_i d\mu &= \sup_n \int f_i^{(n)} d\mu_n = \sup_n \int h_i d(\mu_n \times \nu_n) \\ &= \int h_i d(\mu \times \nu) \end{aligned}$$

(the first equality, by 7.4.5; the second, by (1); the third, by 7.4.5 applied

to the sequence of measures $\mu_n \times \nu_n \uparrow \mu \times \nu$ and the 'constant sequence' of functions $h_i, h_i, h_i \dots$)

claim: $f_i \uparrow \mu$ -a.e.

Let us show, for example, that $f_1 \leq f_2$ a.e. (with respect to μ). Since $f_1^{(n)} \uparrow f_1$ a.e. and $f_2^{(n)} \uparrow f_2$ a.e., there exists a set $E \in \mathcal{S}$ such that $\mu(E) = 0$ and such that, for all $x \in X - E$,

$$f_1(x) = \sup_n f_1^{(n)}(x), \quad f_2(x) = \sup_n f_2^{(n)}(x).$$

Let $x \in X - E$; by (2), we have

$$f_1^{(n)}(x) \leq f_2^{(n)}(x) \quad \text{for all } n,$$

and $f_1(x) \leq f_2(x)$ results on taking supremum over n . Thus $f_1 \leq f_2$ a.e. Similarly $f_i \leq f_{i+1}$ a.e. for each i . Since a countable union of null sets is null, we conclude that $f_i \uparrow$ a.e.

Summarizing, we have $f_i \in \mathcal{L}^1(\mu)$ for all i , $f_i \uparrow \mu$ -a.e., and, citing (4) and the definition of integral (4.4.3), we have

$$(5) \quad \int f_i d\mu = \int h_i d(\mu \times \nu) \uparrow \int h d(\mu \times \nu) < +\infty;$$

by the monotone convergence theorem, there exists $f \in \mathcal{L}^1(\mu)$ such that $f_i \uparrow f$ μ -a.e., and

$$(6) \quad \int f d\mu = \sup_i \int f_i d\mu = \int h d(\mu \times \nu).$$

Let $E \in \mathcal{S}$ with $\mu(E) = 0$ and $f_i \uparrow f$ pointwise on $X - E$. For each i , $f_i^{(n)} \uparrow f_i$ μ -a.e.; choose $E_i \in \mathcal{S}$ with $\mu(E_i) = 0$ and $f_i^{(n)} \uparrow f_i$ on $X - E_i$. Let

$$A = E \cup \bigcup_{i=1}^{\infty} E_i;$$

then $\mu(A) = 0$ and

$$(7) \quad x \in X - A \Rightarrow +\infty > f(x) = \sup_i f_i(x) = \sup_i \left(\sup_n f_i^{(n)}(x) \right) \\ = \sup_i \left(\sup_n \int (h_i)_x d\nu_n \right),$$

in particular,

$$(8) \quad x \in X - A \Rightarrow \int (h_i)_x d\nu_n \leq f(x) < +\infty \quad \text{for all } i \text{ and } n.$$

Fix $x \in X - A$. Fix n . Then

$$\sup_i \int (h_i)_x d\nu_n \leq f(x) < +\infty$$

by (8); since $(h_i)_x \uparrow h_x$ and h_x is measurable, we conclude that $h_x \in \mathcal{L}^1(\nu_n)$ and

$$(9) \quad \int h_x d\nu_n = \sup_i \int (h_i)_x d\nu_n \leq f(x) < +\infty.$$

Now vary n ; since $\nu_n \uparrow \nu$, by Lemma 7.4.5 we conclude that $h_x \in \mathcal{L}^1(\nu)$ and

$$(10) \quad \int h_x d\nu = \sup_n \int h_x d\nu_n.$$

Gathering up all the strands, we have $f \in \mathcal{L}^1(\mu)$, $\mu(A) = 0$ and, for all $x \in X - A$,

$$\begin{aligned} \int h_x d\nu &= \sup_n \int h_x d\nu_n \\ &= \sup_n \left(\sup_i \int (h_i)_x d\nu_n \right) \\ &= \sup_i \left(\sup_n \int (h_i)_x d\nu_n \right) \\ &= f(x) \end{aligned}$$

(the first equality, by (10); the second, by (9); the third, by the 'associativity of sups,' cf. §1.15, Exercise 7; the fourth, by (7)), thus $\iint h d\nu d\mu$ exists and is equal to $\int f d\mu$ —hence, by (6), to $\int h d(\mu \times \nu)$.

(a) \Rightarrow (c): The proof is similar, leading to the equality $\iint h d\mu d\nu = \int h d(\mu \times \nu)$.

(b) \Rightarrow (a): Suppose $\iint h d\nu d\mu$ exists. Let $E \in \mathcal{S}$ and $f \in \mathcal{L}^1(\mu)$ be such that $\mu(E) = 0$ and

$$x \in X - E \quad \Rightarrow \quad h_x \in \mathcal{L}^1(\nu) \quad \text{and} \quad \int h_x d\nu = f(x).$$

Let $x \in X - E$; then $0 \leq (h_i)_x \uparrow h_x \in \mathcal{L}^1(\nu)$, where the $(h_i)_x$ are measurable (7.3.2, (iii)) hence ν -integrable (4.4.20), and

$$\begin{aligned} f(x) &= \int h_x d\nu = \sup_i \int (h_i)_x d\nu \\ &= \sup_i \left(\sup_n \int (h_i)_x d\nu_n \right) \\ &= \sup_i \left(\sup_n f_i^{(n)}(x) \right) \end{aligned}$$

(the first equality, by the choice of E and f ; the second, by the monotone convergence theorem; the third, by Lemma 7.4.5; the fourth by the definition of $f_i^{(n)}$). In particular, $0 \leq f_i^{(n)} \leq f$ μ -a.e. Since f is μ -integrable and $f_i^{(n)}$ is measurable with respect to \mathcal{S} , it follows (by 4.4.20) that

$f_i^{(n)} \in \mathcal{L}^1(\mu)$ and

$$(11) \quad \int f_i^{(n)} d\mu \leq \int f d\mu \quad \text{for all } i \text{ and } n.$$

Since $\mu_n \leq \mu$ and $f_i^{(n)}$ is integrable for both μ_n and μ , it follows that

$$\int f_i^{(n)} d\mu_n \leq \int f_i^{(n)} d\mu \leq \int f d\mu + \infty \quad \text{for all } i \text{ and } n$$

(the first inequality is easily shown by approximating $f_i^{(n)}$ by an increasing sequence of positive simple functions that are integrable with respect to μ , hence with respect to μ_n ; the second, by (11)); thus, citing (1), we have

$$(12) \quad \int h_i d(\mu_n \times \nu_n) \leq \int f d\mu \quad \text{for all } i \text{ and } n.$$

For fixed i , (12) and Lemma 7.4.5 (applied in the context of $\mu_n \times \nu_n \uparrow \mu \times \nu$) yield $h_i \in \mathcal{L}^1(\mu \times \nu)$ (not news—see the first paragraph of the proof) and

$$\int h_i d(\mu \times \nu) = \sup_n \int h_i d(\mu_n \times \nu_n) \leq \int f d\mu < +\infty;$$

thus, the sequence $\int h_i d(\mu \times \nu)$ is bounded, therefore h is $\mu \times \nu$ -integrable (4.4.1).

(c) \Rightarrow (a): Similarly. \diamond

7.4.8. Corollary. (Fubini's theorem) *If μ and ν are σ -finite measures and $h \in \mathcal{L}^1_{\mathbb{C}}(\mu \times \nu)$, then both iterated integrals of h exist and are equal to $\int h d(\mu \times \nu)$.*

Proof. Linearity (cf. 7.3.4, (ii)) and the preceding theorem. \diamond

7.4.9. Example. The condition $h \geq 0$ in the Fubini-Tonelli theorem is essential: in general there exist real-valued measurable functions h on $X \times Y$ such that both iterated integrals exist and are equal, but such that h is not integrable with respect to $\mu \times \nu$.

For example, in the product space $[-1, 1] \times \mathbb{R}$ (each factor equipped with Lebesgue measure), let M be the shaded region (Figure 1) between the curve $y = \frac{|x|}{1-|x|}$ (the upper boundary) and the curve $y = \frac{-|x|}{1-|x|}$ (the lower boundary) and, for $i = \text{I, II, III, IV}$, let M_i be the part of M in quadrant i . Thus, M_1 is the region bounded by the curve $y = \frac{x}{1-x}$ ($0 \leq x < 1$), the x -axis and the line $x = 1$; M_2 is obtained from M_1 by reflecting it in the y -axis; and M_3, M_4 are obtained from M_2, M_1 by reflection in the x -axis.

Write $I = (-1, 1) \times \{0\}$ and define $h : [-1, 1] \times \mathbb{R} \rightarrow \mathbb{R}$ by the formula

$$h(x, y) = \begin{cases} 0 & \text{on } I \\ +1 & \text{on } M_1 - I \\ -1 & \text{on } M_2 - I \\ +1 & \text{on } M_3 - I \\ -1 & \text{on } M_4 - I \\ 0 & \text{on } [-1, 1] \times \mathbb{R} - M. \end{cases}$$

Then $|h| = \varphi_M$ is not integrable with respect to the product measure (because M has infinite measure), therefore neither is h ; however, h is measurable and every section of h is integrable with integral 0, thus both iterated integrals of h exist and are equal to 0.

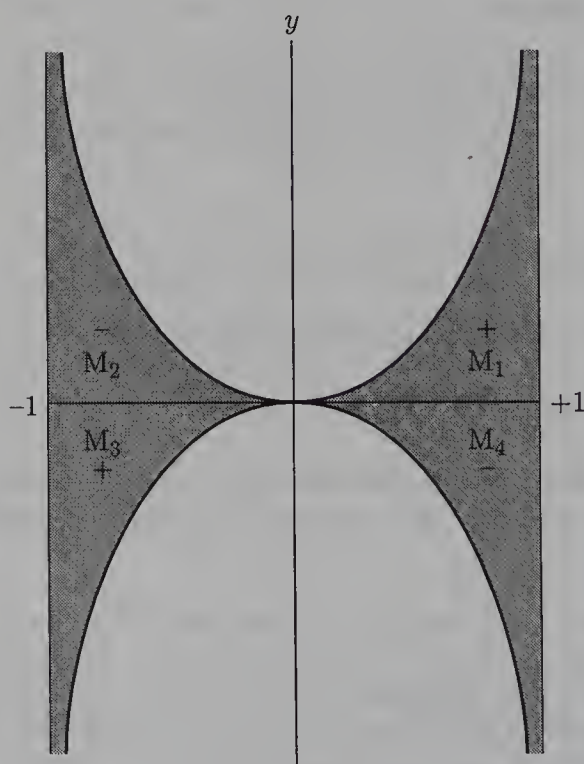


Figure 1

Exercises

1. For an arbitrary (not necessarily σ -finite) measure μ , define contractions as in 7.4.2. In order that μ_A be σ -finite, it is necessary and sufficient that $A = \bigcup_{n=1}^{\infty} E_n$ with $E_n \in \mathcal{S}$ and $\mu(E_n) < +\infty$ for all n .

2. Let (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) be σ -finite measure spaces, let $f \in \mathcal{L}^1_{\mathbb{C}}(\mu)$, $g \in \mathcal{L}^1_{\mathbb{C}}(\nu)$, and define a function $f \otimes g : X \times Y \rightarrow \mathbb{C}$ by the formula

$$(f \otimes g)(x, y) = f(x)g(y) \quad (x \in X, y \in Y).$$

Prove that $f \otimes g \in \mathcal{L}^1(\mu \times \nu)$ and $\int (f \otimes g) d(\mu \times \nu) = (\int f d\mu) (\int g d\nu)$.

{Hint: Reduce, by bilinearity, to the case that $f \geq 0$, $g \geq 0$. After approximating f and g by simple functions, the crux of the matter is that $\varphi_E \otimes \varphi_F = \varphi_{E \times F}$ for every measurable rectangle $E \times F$.}

3. Let (X, \mathcal{S}, μ) and (Y, \mathcal{T}, ν) be σ -finite measure spaces, $\pi = \mu \times \nu$ the product measure on $\mathcal{S} \times \mathcal{T}$.

(i) The following conditions on a set $M \in \mathcal{S} \times \mathcal{T}$ are equivalent: (a) $\pi(M) = 0$; (b) $\nu(M_x) = 0$ for μ -almost every $x \in X$; (c) $\mu(M^y) = 0$ for ν -almost every $y \in Y$.

(ii) Let $h, k : X \times Y \rightarrow \mathbb{C}$ be functions such that $h = k$ a.e. (with respect to π). If $\iint h d\nu d\mu$ exists and k_x is measurable for μ -almost every x , then $\iint k d\nu d\mu$ exists and is equal to $\iint h d\nu d\mu$. When ν is complete (§2.4, Exercise 8, the measurability assumption on k can be omitted.

(iii) The condition $h \geq 0$ in the Fubini–Tonelli theorem can be relaxed to $h \geq 0$ a.e.

(iv) (*Cavalieri's Principle*)¹ If M and N are sets in $\mathcal{S} \times \mathcal{T}$ such that $\nu(M_x) = \nu(N_x)$ for μ -almost every $x \in X$, then $\pi(M) = \pi(N)$.

(v) Specialize to $(Y, \mathcal{T}, \nu) = (\mathbb{R}, \mathcal{B}, \lambda)$, where λ is the restriction of Lebesgue measure to the σ -algebra \mathcal{B} of Borel sets in \mathbb{R} ; thus $\pi = \mu \times \lambda$. Suppose $f \in \mathcal{L}^1_{\mathbb{R}}(\mu)$, $f \geq 0$, and let

$$M = \{(x, y) : x \in X, 0 \leq y < f(x)\},$$

$$N = \{(x, y) : x \in X, 0 \leq y \leq f(x)\}.$$

We know that $\pi(M) = \int f d\mu$ (§7.2, Exercise 6); infer that $\pi(N) = \pi(M)$, consequently the graph $N - M = \{(x, f(x)) : x \in X\}$ of f has measure 0.

{Hint: (i) Consider $h = \varphi_M$ in the Fubini–Tonelli theorem. (ii) If $M \in \mathcal{S} \times \mathcal{T}$ with $\pi(M) = 0$ and $h = k$ on $X \times Y - M$, observe that $h_x = k_x$ on $Y - M_x$ and contemplate (i); for the case that ν is complete, see §4.2, Exercise 4. (iv) It suffices to prove that if $\pi(M)$ is finite then $\pi(N) = \pi(M)$.}

¹P.R. Halmos, *Measure theory* [Van Nostrand, New York, 1950; reprinted Springer-Verlag, New York, 1974], p. 149, Exercise (5).

CHAPTER 8

The Differential Equation $y' = f(x, y)$

§8.1. Equicontinuity, Ascoli's theorem

§8.2. Picard's existence theorem for $y' = f(x, y)$

§8.3. Peano's existence theorem for $y' = f(x, y)$

Our objective in this chapter is to solve the differential equation $y' = f(x, y)$ for a suitable class of continuous real-valued functions f of two real variables x and y , where a "solution" is understood to be a real-valued (continuously differentiable) function $\varphi : I \rightarrow \mathbb{R}$, defined on a suitable interval I , such that $\varphi'(x) = f(x, \varphi(x))$ for all $x \in I$.

The strategy is to prove the theorem first for a special class of continuous functions f (§8.2) that contains the polynomial functions, then extend the theorem to the case of general f (§8.3) by approximating f by a sequence (f_n) of functions in the special class just mentioned (in fact, polynomial functions), constructing a corresponding sequence of functions φ_n satisfying $\varphi'_n(x) = f_n(x, \varphi_n(x))$ by the special case already proved, then passing to the limit on n ; to preserve differentiability under the passage to the limit, we had better exhibit f as the *uniform* limit of the f_n (cf. 6.2.27), the uniform approximability being possible by the Weierstrass theorem. Now, the sequence (φ_n) corresponding to the f_n need not be uniformly convergent, but it is sufficient that it have a uniformly convergent subsequence (φ_{n_k}) . We are thus led to contemplate sets of functions that are compact for the uniform metric topology (cf. 6.1.23); the mission of Ascoli's theorem (§8.1) is to characterize such sets (by means of criteria that are verifiable in the application at hand).

8.1. Equicontinuity, Ascoli's Theorem

Notations fixed throughout this section: (X, d) and (Y, ρ) are metric spaces, and we are interested in sets of continuous functions $X \rightarrow Y$, that is, in subsets \mathcal{E} of $\mathcal{C}(X, Y)$ (6.2.20). If X is compact, then every continuous function $X \rightarrow Y$ is bounded (6.2.11) and the formula

$$D(f, g) = \sup_{x \in X} \rho(f(x), g(x))$$

defines a metric D on $\mathcal{C}(X, Y)$ such that

$$D(f_n, f) \rightarrow 0 \Leftrightarrow f_n \rightarrow f \text{ uniformly on } X$$

(see 6.2.15); one calls the topology on $\mathcal{C}(X, Y)$ derived from D the **topology of uniform convergence**. The most memorable result (8.1.15) characterizes the compact subsets of $\mathcal{C}(X, Y)$ for the topology of uniform convergence, assuming that X is compact and (Y, ρ) is complete; the following definitions introduce the concepts that are needed in the characterization.

8.1.1. *Definition.* Let T be a set, $\mathcal{S} \subset \mathcal{F}(T, Y)$ a set of functions on T with values in the metric space (Y, ρ) . We say that \mathcal{S} is **totally bounded at a point** $t \in T$ if the set

$$\mathcal{S}(t) = \{f(t) : f \in \mathcal{S}\}$$

is a totally bounded metric subspace of (Y, ρ) , that is, admits an ϵ -net for every $\epsilon > 0$ (6.1.13); an equivalent condition is that every sequence in $\mathcal{S}(t)$ has a Cauchy subsequence (6.1.24). We say that \mathcal{S} is **pointwise totally bounded** (on T) if it is totally bounded at every point of T (a condition automatically satisfied when (Y, ρ) is itself a totally bounded metric space).

8.1.2. *Definition.* Let T be a topological space, $\mathcal{E} \subset \mathcal{F}(T, Y)$ a set of functions on T with values in the metric space (Y, ρ) . We say that \mathcal{E} is **equicontinuous at a point** $t \in T$ if, for every $\epsilon > 0$, there exists a neighborhood V of t such that

$$(*) \quad f(V) \subset U_\epsilon(f(t)) \quad \text{for all } f \in \mathcal{E}$$

(where $U_\epsilon(f(t))$ denotes the open ball of radius ϵ centered at $f(t)$), that is,

$$t' \in V \Rightarrow \rho(f(t'), f(t)) < \epsilon \quad \text{for all } f \in \mathcal{E}.$$

8.1.3. *Remark.* The condition $(*)$ is equivalent to

$$V \subset f^{-1}(U_\epsilon(f(t))) \quad \text{for all } f \in \mathcal{E},$$

that is, to

$$V \subset \bigcap_{f \in \mathcal{E}} f^{-1}(U_\epsilon(f(t)));$$

since the set on the right is a neighborhood of t if and only if it contains a neighborhood of t , the equicontinuity of \mathcal{E} at t is equivalent to the

condition

$$(\forall \epsilon > 0) \bigcap_{f \in \mathcal{E}} f^{-1}(U_\epsilon(f(t))) \text{ is a neighborhood of } t.$$

In particular, it is clear that if \mathcal{E} is equicontinuous at t , then every $f \in \mathcal{E}$ is continuous at t (3.4.2).

8.1.4. *Definition.* With notations $\mathcal{E} \subset \mathcal{F}(T, Y)$ as in 8.1.2, we say that \mathcal{E} is **equicontinuous** (on T) if it is equicontinuous at every point $t \in T$.

8.1.5. *Remark.* If $\mathcal{E} \subset \mathcal{F}(T, Y)$ is equicontinuous on T , then every $f \in \mathcal{E}$ is continuous on T (cf. 8.1.3).

8.1.6. *Definition.* Let $\mathcal{E} \subset \mathcal{F}(X, Y)$ be a set of functions on the metric space (X, d) with values in the metric space (Y, ρ) . We say that \mathcal{E} is **equi-uniformly continuous** (on X) if, for every $\epsilon > 0$, there exists a $\delta > 0$ such that

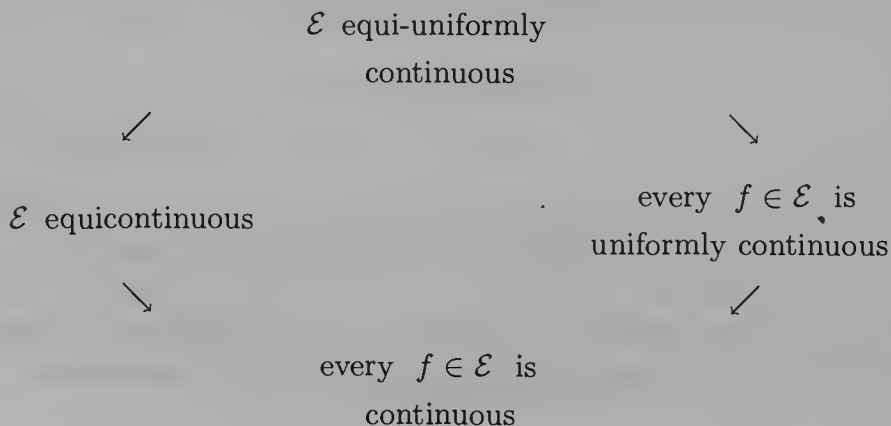
$$(**) \quad d(x, x') \leq \delta \Rightarrow \rho(f(x), f(x')) \leq \epsilon \text{ for all } f \in \mathcal{E}.$$

(The term 'uniformly equicontinuous' is also used.)

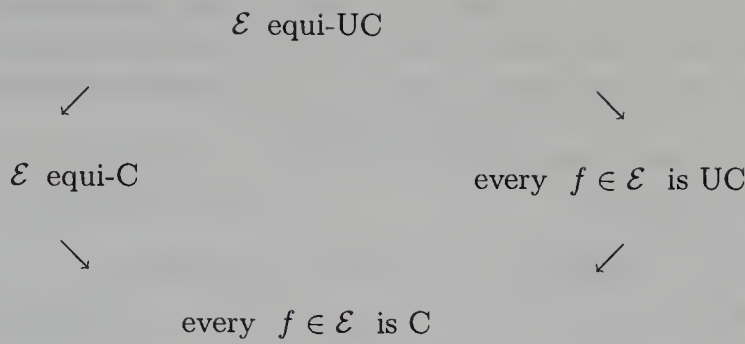
8.1.7. *Remark.* With notations as in 8.1.6, to say that a function $f : X \rightarrow Y$ is uniformly continuous means (6.3.2) that, for every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$d(x, x') \leq \delta \Rightarrow \rho(f(x), f(x')) \leq \epsilon;$$

thus, to say that \mathcal{E} is equi-uniformly continuous on X means that, for every $\epsilon > 0$, there exists a $\delta > 0$ that 'works' simultaneously for all functions $f \in \mathcal{E}$. Thus, for a set $\mathcal{E} \subset \mathcal{F}(X, Y)$, we have the diagram of implications



or, more succinctly,



The central result to be proved is as follows:

8.1.8. Theorem. (Ascoli's theorem)¹ *Let $f_n : X \rightarrow Y$ ($n = 1, 2, 3, \dots$) be a sequence of functions, where (X, d) is a compact metric space and (Y, ρ) is a metric space. If the set*

$$\mathcal{E} = \{f_n : n = 1, 2, 3, \dots\} \subset \mathcal{C}(X, Y)$$

is equicontinuous and pointwise totally bounded, then there exists a subsequence (f_{n_k}) that is uniformly Cauchy.

The proof of Ascoli's theorem is arranged in a series of lemmas. Although the concept of equi-uniform continuity does not appear explicitly in the statement of the theorem, it arises in the proof and is, in fact, equivalent to equicontinuity in the presence of the compactness of X (Lemma 8.1.10 below).

8.1.9. Lemma. *If (X, d) is a compact metric space and \mathcal{U} is any open covering of X , then there exists a real number $\tau > 0$ such that*

$$\left. \begin{array}{l} A \subset X \\ \text{diam } A \leq \tau \end{array} \right\} \Rightarrow A \subset U \text{ for some } U \in \mathcal{U},$$

*that is, every subset of X of diameter $\leq \tau$ is entirely contained in at least one of the sets of the covering. (Such a number τ is called a **Lebesgue number** for the covering.)*

Proof. Assume to the contrary that every $\tau > 0$ fails to have the desired property. Then, for every positive integer n , the number $1/n$ fails to have the property, so there exists a subset A_n of X such that $\text{diam } A_n \leq 1/n$ but no one set of \mathcal{U} contains A_n , that is,

(*) $(\forall U \in \mathcal{U}) \quad A_n \not\subset U.$

¹Giulio Ascoli (1843–1896). For an interesting historical note, see N. Bourbaki, *General topology. Vol. II (Chs. 5-10)* [Springer-Verlag, New York, 1988], p. 347.

In particular, every A_n is nonempty; for each n , choose a point $x_n \in A_n$ such that $x_n \notin U$. Since X is compact, the sequence (x_n) has a convergent subsequence (6.1.11), say $x_{n_k} \rightarrow x$. Since \mathcal{U} covers X , there exists a set $U \in \mathcal{U}$ such that $x \in U$. A contradiction to $(*)$ will be reached by showing that $A_{n_k} \subset U$ for some index k .

Since U is open, there exists an $\epsilon > 0$ such that $U_\epsilon(x) \subset U$. Choose an index k such that

$$d(x_{n_k}, x) < \epsilon/2 \quad \text{and} \quad 1/n_k < \epsilon/2;$$

it will suffice to show that $A_{n_k} \subset U_\epsilon(x)$. Let $y \in A_{n_k}$. By the triangle inequality,

$$d(y, x) \leq d(y, x_{n_k}) + d(x_{n_k}, x),$$

where $d(y, x_{n_k}) \leq 1/n_k$ (because $\text{diam } A_{n_k} \leq 1/n_k$), therefore

$$d(y, x) \leq 1/n_k + d(x_{n_k}, x) < \epsilon/2 + \epsilon/2 = \epsilon$$

by the choice of k , whence $y \in U_\epsilon(x)$. \diamond

8.1.10. Lemma. Let $\mathcal{E} \subset \mathcal{F}(X, Y)$ be a set of functions from X to Y , where (X, d) and (Y, ρ) are the given metric spaces and X is compact. Then

$$\mathcal{E} \text{ is equicontinuous} \Leftrightarrow \mathcal{E} \text{ is equi-uniformly continuous.}$$

Proof. \Leftarrow : Noted in 8.1.7.

\Rightarrow : Let $\epsilon > 0$. For each $x \in X$, choose $\delta_x > 0$ so that

$$f(U_{\delta_x}(x)) \subset U_{\epsilon/2}(f(x)) \quad \text{for all } f \in \mathcal{E}.$$

The family $(U_{\delta_x}(x))_{x \in X}$ is an open covering of X ; let $\delta > 0$ be a Lebesgue number for the covering (8.1.9). If $x, x' \in X$, then

$$\begin{aligned} d(x, x') < \delta &\Rightarrow \text{diam}\{x, x'\} < \delta \\ &\Rightarrow \{x, x'\} \subset U_{\delta_x}(z) \quad \text{for some } z \in X \\ &\Rightarrow f(\{x, x'\}) \subset f(U_{\delta_x}(z)) \subset U_{\epsilon/2}(f(z)) \quad \text{for all } f \in \mathcal{E} \\ &\Rightarrow \rho(f(x), f(x')) < \epsilon \quad \text{for all } f \in \mathcal{E}, \end{aligned}$$

thus \mathcal{E} is equi-uniformly continuous (8.1.6). \diamond

8.1.11. Lemma. Let (X, d) and (Y, ρ) be the given metric spaces and suppose that (X, d) is **totally bounded**. Let (f_n) be an equi-uniformly continuous sequence in $\mathcal{C}(X, Y)$ and suppose there exists a dense subset A of X such that $(f_n(x))$ is Cauchy for each $x \in A$. Then (f_n) is uniformly Cauchy on X .

Proof. We observe first that if $f : X \rightarrow Y$ is uniformly continuous, then $f(X)$ is a totally bounded subset of Y . For, let (y_n) be any sequence in $f(X)$ and, for each n , choose $x_n \in X$ so that $f(x_n) = y_n$. Since X

is totally bounded, the sequence (x_n) has a Cauchy subsequence (x_{n_k}) (6.1.24), and since f is uniformly continuous it follows that $(f(x_{n_k})) = (y_{n_k})$ is Cauchy in Y (6.3.4). Thus, every sequence in $f(X)$ has a Cauchy subsequence, consequently $f(X)$ is totally bounded (6.1.24). In particular,

$$\text{diam } f(X) < +\infty,$$

that is, f is bounded (6.2.9). Thus,

$$f : X \rightarrow Y \text{ uniformly continuous} \Rightarrow f \in \mathcal{C}(X, Y) \cap \mathcal{B}(X, Y),$$

so that the metric

$$D(f, g) = \sup_{x \in X} \rho(f(x), g(x))$$

on $\mathcal{B}(X, Y)$ is defined for all pairs f, g of uniformly continuous functions $X \rightarrow Y$; in other words, D defines a metric on the set $\mathcal{C}_u(X, Y)$ of all uniformly continuous functions $X \rightarrow Y$, convergence relative to D signifying uniform convergence (6.2.15).

With (f_n) as in the statement of the lemma, we are to show that $D(f_m, f_n) \rightarrow 0$ as $m, n \rightarrow \infty$. Given any $\epsilon > 0$, we seek an index N such that

$$m, n \geq N \Rightarrow D(f_m, f_n) \leq \epsilon.$$

By the assumed equi-uniform continuity of the sequence, there exists a $\delta > 0$ such that

$$(1) \quad d(x, x') < \delta \Rightarrow \rho(f_n(x), f_n(x')) < \epsilon/3 \text{ for all } n.$$

Let x_1, \dots, x_r be a $\delta/2$ -net in the totally bounded space X ; by the density of A , there exist points a_1, \dots, a_r in A such that

$$d(a_i, x_i) < \delta/2 \text{ for } i = 1, \dots, r,$$

and it follows from the triangle inequality that

$$(2) \quad a_1, \dots, a_r \text{ is a } \delta\text{-net in } X.$$

By the hypothesis on A , we know that each of the finitely many sequences $(f_n(a_i))$ ($i = 1, \dots, r$) is Cauchy in Y , therefore there exists an index N such that

$$(3) \quad m, n \geq N \Rightarrow \rho(f_m(a_i), f_n(a_i)) < \epsilon/3 \text{ for } i = 1, \dots, r.$$

Fix a pair of indices $m, n \geq N$ and fix a point $x \in X$; we need only show that $\rho(f_m(x), f_n(x)) < \epsilon$.

By (2), there exists an index j ($1 \leq j \leq r$) such that

$$(4) \quad d(x, a_j) < \delta.$$

Then

$$\rho(f_m(x), f_n(x)) \leq \rho(f_m(x), f_m(a_j)) + \rho(f_m(a_j), f_n(a_j)) + \rho(f_n(a_j), f_n(x)),$$

where the middle term on the right side of the inequality is $< \epsilon/3$ by (3), while the first and third terms on the right are $< \epsilon/3$ by (4) and (1), thus $\rho(f_m(x), f_n(x)) < \epsilon$ as desired. \diamond

8.1.12. Lemma. *If A is a countable set and $g_n : A \rightarrow Y$ is a pointwise totally bounded sequence of functions on A with values in the metric space (Y, ρ) , then there exists a subsequence of (g_n) that is pointwise Cauchy on A .*

Proof. Orientation: (i) When A is a singleton, the assertion is immediate from Theorem 6.1.24 (whence it is true for A finite, by an easy iteration); (ii) every totally bounded (hence every compact) metric space has a countable dense subset, thus Lemmas 8.1.12 and 8.1.11 effectively reduce the proof of Ascoli's theorem to a theorem about functions on a countable totally bounded metric space.

The method of proof is the so-called 'diagonal method'. Say

$$A = \{a_1, a_2, a_3, \dots\}.$$

Apply the functions g_n to the point a_1 : by hypothesis, the set

$$S = \{g_n(a_1) : n \in \mathbb{P}\}$$

is totally bounded, therefore $(g_n(a_1))$ is a sequence in the totally bounded metric space $(S, \rho|_{S \times S})$; by Theorem 6.1.24, there exists a subsequence (g_n^1) of (g_n) such that

$$(1) \quad (g_n^1(a_1)) \text{ is a Cauchy sequence in } Y.$$

Say $g_n^1 = g_{i_n}$, where $i_1 < i_2 < i_3 < \dots$. By an obvious induction, $i_k \geq k$ for all k ;

$$g_1, \quad g_2, \quad g_3, \quad \dots, \quad g_k, \quad \dots$$

$$g_1^1, \quad g_2^1, \quad g_3^1, \quad \dots, \quad g_k^1, \quad \dots$$

so to speak, the k th term of (g_n^1) is 'at least as far along' in the sequence (g_n) as the k th term of (g_n) . Moreover, if $k > j$ then g_k^1 is strictly to the right of g_j^1 in the sequence g_1, g_2, g_3, \dots (because $i_k > i_j$).

Now apply the sequence (g_n^1) to the point a_2 and repeat the preceding argument: there exists a subsequence (g_n^2) of (g_n^1) , say $g_n^2 = g_{j_n}^1$ ($j_1 < j_2 < j_3 < \dots$), such that

$$(2) \quad (g_n^2(a_2)) \text{ is a Cauchy sequence in } Y.$$

As before, the k th term of (g_n^2) is at least as far along in the sequence (g_n^1)

as is g_k^1 .

$$\begin{array}{ccccccc} g_1, & g_2, & g_3, & \dots, & g_k, & \dots \\ g_1^1, & g_2^1, & g_3^1, & \dots, & g_k^1, & \dots \\ g_1^2, & g_2^2, & g_3^2, & \dots, & g_k^2, & \dots \end{array}$$

Moreover, if $k > j$ then g_k^2 is strictly to the right of g_j^2 in the sequence (g_n^1) , hence (by the preceding paragraph) in the sequence (g_n) . In particular, g_2^2 is strictly to the right of g_1^1 in the sequence (g_n^1) (because $j_2 \geq 2 > 1$), hence in the sequence (g_n) . Also, $(g_n^2(a_1))$ is a subsequence of $(g_n^1(a_1))$, therefore

$$(g_n^2(a_1)) \text{ is a Cauchy sequence in } Y.$$

Continuing recursively, we construct a tableau of functions

$$\begin{array}{cccc} g_1^1, & g_2^1, & g_3^1, & \dots \\ g_1^2, & g_2^2, & g_3^2, & \dots \\ g_1^3, & g_2^3, & g_3^3, & \dots \end{array}$$

such that each row is a subsequence of the preceding row (hence of all preceding rows, and of the original sequence g_1, g_2, g_3, \dots), and such that for each k ,

$$(g_n^k(a_j)) \text{ is a Cauchy sequence in } Y, \text{ for } j = 1, \dots, k,$$

that is, if the k th row of the tableau is applied to any of the elements a_1, \dots, a_k , the result is a Cauchy sequence in Y . Moreover, the element g_k^k of the k th row is strictly to the right of g_{k-1}^{k-1} in the $(k-1)$ th row (hence in the original sequence g_1, g_2, g_3, \dots).

We now look at the sequence

$$g_1^1, \quad g_2^2, \quad g_3^3, \quad \dots,$$

the *diagonal* of the tableau. From the foregoing remarks, it is clear that this is a subsequence of the original sequence (g_n) . Moreover, for each k , the ‘truncated diagonal’

$$g_k^k, \quad g_{k+1}^{k+1}, \quad g_{k+2}^{k+2}, \quad \dots$$

is a subsequence of the k th row, thus the sequence

$$g_k^k(a_k), \quad g_{k+1}^{k+1}(a_k), \quad g_{k+2}^{k+2}(a_k), \quad \dots$$

is Cauchy in Y , therefore so is the sequence $(g_n^n(a_k))$. In other words, the diagonal sequence (g_n^n) is a subsequence of (g_n) that is pointwise Cauchy on the set A . \diamond

We are now ready to prove Ascoli's theorem.

Proof of Theorem 8.1.8: By assumption X is compact, therefore totally bounded (6.1.12, 6.1.13), therefore separable (6.1.19); let A be a countable dense subset of X . The sequence (f_n) , assumed to be pointwise totally bounded on X , is in particular pointwise totally bounded on A ; applying the preceding lemma to the sequence of restrictions $g_n = f_n|_A$, we obtain a subsequence (f_{n_k}) that is pointwise Cauchy on A . Since the f_n are equicontinuous by hypothesis, so are the f_{n_k} ; by Lemma 8.1.10, the f_{n_k} are equi-uniformly continuous, so it follows from Lemma 8.1.11 that the sequence f_{n_k} is uniformly Cauchy on X . \diamond

8.1.13. Corollary. *As in Ascoli's theorem, let (X, d) be a compact metric space and (Y, ρ) a metric space. If $\mathcal{E} \subset \mathcal{C}(X, Y)$ is equicontinuous and pointwise totally bounded, then \mathcal{E} is totally bounded for the sup-metric.*

Proof. By Ascoli's theorem, every sequence in \mathcal{E} has a uniformly Cauchy subsequence, therefore \mathcal{E} is totally bounded for the sup-metric by Theorem 6.1.24. \diamond

8.1.14. Corollary. *Let (X, d) be a compact metric space and (Y, ρ) a complete metric space. The following conditions on a set $\mathcal{E} \subset \mathcal{C}(X, Y)$ are equivalent:*

- (a) \mathcal{E} is totally bounded (for the sup-metric);
- (b) $\overline{\mathcal{E}}$ is compact (for the topology of uniform convergence, where the bar denotes closure relative to that topology);
- (c) \mathcal{E} is pointwise totally bounded and equicontinuous (hence equi-uniformly continuous).

Proof. (c) \Rightarrow (a): This is Corollary 8.1.13.

(a) \Rightarrow (b): Given any $\epsilon > 0$, let f_1, \dots, f_r be an $\epsilon/2$ -net for \mathcal{E} . If $f \in \overline{\mathcal{E}}$ then f is within $\epsilon/2$ of some point of \mathcal{E} , hence within ϵ of some f_i , thus f_1, \dots, f_r is an ϵ -net for $\overline{\mathcal{E}}$. Thus, the total boundedness of \mathcal{E} implies that of $\overline{\mathcal{E}}$ (and the argument is valid in any metric space). Since $\mathcal{C}(X, Y)$ is complete for the sup-metric (6.2.22), its closed subset $\overline{\mathcal{E}}$ is also complete (6.1.29). Thus, $\overline{\mathcal{E}}$ is complete and totally bounded, hence compact (6.1.26).

(b) \Rightarrow (c): *Proof that \mathcal{E} is pointwise totally bounded.* Let $x \in X$. The mapping $\mathcal{C}(X, Y) \rightarrow Y$ defined by $f \mapsto f(x)$ is continuous (because uniform convergence implies pointwise convergence), consequently the image $\overline{\mathcal{E}}(x) = \{f(x) : f \in \overline{\mathcal{E}}\}$ of the compact set $\overline{\mathcal{E}}$ is compact in Y (6.1.27), therefore is totally bounded; it follows easily that its subset $\mathcal{E}(x)$ is also totally bounded (cf. the proof of (2) of 8.1.11).

Proof that \mathcal{E} is equicontinuous: In view of Lemma 8.1.10, it is the same to show that \mathcal{E} is equi-uniformly continuous. Let $\epsilon > 0$. We seek a $\delta > 0$ such that

$$d(x, x') < \delta \quad \Rightarrow \quad \rho(f(x), f(x')) < \epsilon \quad \text{for all } f \in \mathcal{E}.$$

Since $\overline{\mathcal{E}}$ is compact, hence totally bounded, it follows that \mathcal{E} is totally bounded; let f_1, \dots, f_r be an $\epsilon/3$ -net for \mathcal{E} . For each index i , f_i is continuous, therefore uniformly continuous (6.3.7), so there exists a $\delta_i > 0$ such that

$$d(x, x') < \delta_i \Rightarrow \rho(f_i(x), f_i(x')) < \epsilon/3;$$

writing $\delta = \min\{\delta_1, \dots, \delta_r\}$, we have

$$(1) \quad d(x, x') < \delta \Rightarrow \rho(f_i(x), f_i(x')) < \epsilon/3 \text{ for } i = 1, \dots, r.$$

Fix $f \in \mathcal{E}$ and let $x, x' \in X$ with $d(x, x') < \delta$; it will suffice to show that $\rho(f(x), f(x')) < \epsilon$. Choose an index j such that $D(f, f_j) < \epsilon/3$ (D the sup-metric), possible since f_1, \dots, f_r is an $\epsilon/3$ -net for \mathcal{E} . Then

$$(2) \quad \rho(f(z), f_j(z)) < \epsilon/3 \text{ for all } z \in X.$$

Thus

$$\rho(f(x), f(x')) \leq \rho(f(x), f_j(x)) + \rho(f_j(x), f_j(x')) + \rho(f_j(x'), f(x')),$$

where the first and third terms on the right side are $< \epsilon/3$ by (2), and the middle term is $< \epsilon/3$ by (1), consequently $\rho(f(x), f(x')) < \epsilon/3 + \epsilon/3 + \epsilon/3$. \diamond

The most memorable consequence of Ascoli's theorem:

8.1.15. Corollary. *With X and Y as in the preceding corollary (X compact and Y complete), the following conditions on a set $\mathcal{E} \subset \mathcal{C}(X, Y)$ are equivalent:*

- (a) \mathcal{E} is compact;
- (b) \mathcal{E} is closed, pointwise totally bounded, and equicontinuous.

Proof. (a) \Rightarrow (b): Since \mathcal{E} is a compact subset of $\mathcal{C}(X, Y)$ (for the topology of uniform convergence), it is both totally bounded and complete for the sup-metric (6.1.26), hence it is a closed subset of $\mathcal{C}(X, Y)$ (6.1.29; more generally, see Exercise 1). By the preceding corollary, \mathcal{E} is pointwise totally bounded and equicontinuous.

(b) \Rightarrow (a): By the preceding corollary, $\overline{\mathcal{E}}$ is compact; by hypothesis \mathcal{E} is a closed set, so $\mathcal{E} = \overline{\mathcal{E}}$ (3.3.16). \diamond

Exercises

1. (i) In a quasicompact topological space (6.1.3), every closed subset is quasicompact.

(ii) In a separated topological space (6.1.6), every quasicompact subset is compact and closed.²

² Cf. J. Dixmier, *General topology* [Springer-Verlag, New York, 1984], p. 39, Theorem 4.2.7.

(iii) If X is quasicompact, Y is separated and $f : X \rightarrow Y$ is a continuous bijection, then f is a homeomorphism.

{Hint: (i) This is §6.1, Exercise 11. (ii) Let A be a quasicompact subset of the separated space X . The compactness of A is noted in 6.1.9, (1). To show that A is closed, assuming $x \notin A$ it suffices to show that $x \notin \overline{A}$, and for this it is enough to find a neighborhood V of x that is disjoint from A . For each $a \in A$ there exist open sets U_a, V_a such that $a \in U_a$, $x \in V_a$ and $U_a \cap V_a = \emptyset$.

(iii) It suffices to show that if A is a closed set in X then $f(A)$ is closed in Y . Cf. Theorem 6.1.27.}

2. Infer the equivalence (a) \Leftrightarrow (b) in Corollary 8.1.14 from Corollary 6.3.16.

{Hint: Corollary 6.2.22, (2).}

8.2. Picard's Existence Theorem for $y' = f(x, y)$

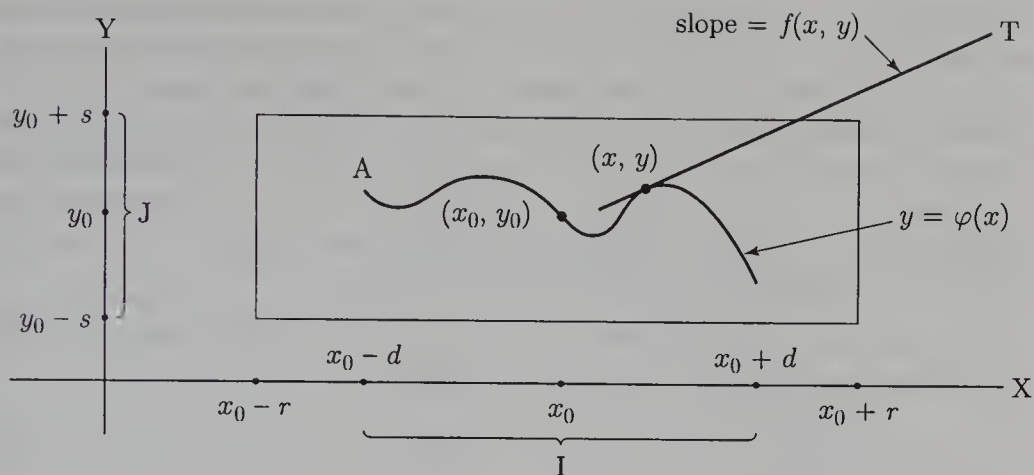
The following motivational remarks can be omitted by the reader who wishes to go straight to the theorem in question (8.2.1).

Given a real-valued function f of two real variables, we seek a real-valued function φ of one real variable such that $\varphi'(x) = f(x, \varphi(x))$ for all x (in the domain of φ), that is, abbreviating $y = \varphi(x)$, such that $y' = f(x, y)$ for all x . It would be nice if φ were defined and differentiable (in particular, continuous) on an interval of x 's.

Let's start over. Let (x_0, y_0) be an interior point of a subset A of the cartesian plane \mathbb{R}^2 , and let $f : A \rightarrow \mathbb{R}$ be a real-valued function on A . We seek a real-valued function $\varphi : I \rightarrow \mathbb{R}$ defined on an interval I containing x_0 as an internal point, such that (i) $\varphi(x_0) = y_0$ and (ii) $\varphi'(x) = f(x, \varphi(x))$ for all $x \in I$. This will require, in particular, that f be defined on the graph of φ , that is, that $\{(x, \varphi(x)) : x \in I\} \subset A$; it would suffice, for example, to make sure that $I \times \varphi(I) \subset A$. Since A is a neighborhood of (x_0, y_0) , it contains a rectangle centered at (x_0, y_0) with sides parallel to the coordinate axes; dropping down to such a rectangle, we can suppose for simplicity of notation that

$$A = [x_0 - r, x_0 + r] \times [y_0 - s, y_0 + s]$$

for suitable $r > 0$, $s > 0$. We seek a subinterval I of $[x_0 - r, x_0 + r]$, say of the form $I = [x_0 - d, x_0 + d]$ with $0 < d \leq r$, and a function $\varphi : I \rightarrow \mathbb{R}$ satisfying the conditions (i) and (ii). Writing $J = [y_0 - s, y_0 + s]$, we can assure that A contains the graph of φ by making sure that $\varphi(I) \subset J$, as this will imply that $I \times \varphi(I) \subset I \times J \subset A$. The picture is as follows:



Assuming $A = [x_0 - r, x_0 + r] \times J$ is the indicated rectangle with center (x_0, y_0) , and given a 'field of slopes' f defined on A , our problem is to thread a curve through (x_0, y_0) in such a way that at every point (x, y) on the curve, the slope of the tangent line T to the curve is the given value $f(x, y)$ of f . It sounds like an impossible task; we start at (x_0, y_0) and ... then what? Any idea for proceeding will require some sort of hypothesis on f ; conversely, a proposed hypothesis on f could give us an idea for proceeding.

Suppose f were continuous on A . Then a solution function φ , satisfying $\varphi'(x) = f(x, \varphi(x))$, would be *continuously differentiable* (the formula imposes continuity on φ'), and that reminds us of the fundamental theorem of calculus:¹

$$\varphi(x) = \int_{x_0}^x \varphi'(t) dt + \varphi(x_0) \quad \text{for all } x \in I,$$

in other words,

$$\varphi(x) = y_0 + \int_{x_0}^x f(t, \varphi(t)) dt \quad (x \in I).$$

The latter formula, expressing φ in terms of itself, suggests trying to construct φ by *successive approximations*: start with a rough guess φ_0 (for instance, the constant function $\varphi_0(x) \equiv y_0$), define

$$\varphi_1(x) = y_0 + \int_{x_0}^x f(t, \varphi_0(t)) dt \quad (x \in I),$$

and, recursively,

$$\varphi_{n+1}(x) = y_0 + \int_{x_0}^x f(t, \varphi_n(t)) dt \quad (x \in I),$$

¹ First course, p. 151, Theorem 9.4.6.

then hope that the sequence (φ_n) converges, in a suitable sense, to a solution function φ (perhaps at the expense of imposing further conditions on f). Along the way, we will have to ensure that the graph of φ_n is contained in A so that φ_{n+1} can be defined. The facts relevant to Picard's theorem: one can take d sufficiently small that $\varphi_n(I) \subset J$ at each step of the construction, and the uniform convergence of (φ_n) can be assured by requiring that f satisfy a Lipschitz condition with respect to the second variable y . It is time to get down to brass tacks:

8.2.1. Theorem. (E. Picard)² Let $f : A \rightarrow \mathbb{R}$ be a continuous real-valued function defined on a compact rectangle

$$A = [x_0 - r, x_0 + r] \times [y_0 - s, y_0 + s] \quad (r > 0, s > 0)$$

in the Euclidean plane \mathbb{R}^2 . Let $M > 0$ be a bound for f on A ,

$$|f(x, y)| \leq M \quad \text{for all } (x, y) \in A,$$

let $d = \min\{r, s/M\}$, and write

$$I = [x_0 - d, x_0 + d], \quad J = [y_0 - s, y_0 + s].$$

Assume, in addition, that there exists a constant $K > 0$ such that

$$(*) \quad |f(x, y_1) - f(x, y_2)| \leq K|y_1 - y_2| \quad \text{for all } (x, y_1), (x, y_2) \in A.$$

Then, there exists a unique continuously differentiable function $\varphi : I \rightarrow \mathbb{R}$ such that

$$(**) \quad \varphi(x_0) = y_0, \quad \varphi(I) \subset J, \quad \text{and} \quad \varphi'(x) = f(x, \varphi(x)) \quad \text{for all } x \in I$$

(where, at the endpoints of I , φ' denotes the one-sided derivative).

(In the next section, the existence of φ will be proved without requiring the Lipschitz condition $(*)$ (Peano's theorem), at the cost of dropping uniqueness in the conclusion.

Proof. Existence: The proof is by "successive approximations". We shall construct recursively a sequence of functions $\varphi_n : I \rightarrow \mathbb{R}$ ($n = 0, 1, 2, 3, \dots$), with $\varphi_n(I) \subset J$ for all n , satisfying the conditions

$$\varphi_n(x) = y_0 + \int_{x_0}^x f(t, \varphi_{n-1}(t)) dt \quad (x \in I, n = 1, 2, 3, \dots),$$

such that, thanks to the Lipschitz condition $(*)$, (φ_n) converges uniformly to a function φ satisfying $(**)$. It is important for the proof of Peano's

²Émile Picard (1856–1941). For a generalization of the theorem with f a vector-valued function defined on a subset A of a Euclidean space \mathbb{R}^k , see G. Birkhoff and G.-C. Rota, *Ordinary differential equations* [3rd edn., Wiley, New York, 1978], p. 152, Theorem 6.

theorem in the next section that the Lipschitz constant K does not figure in the definition of the interval I .

Define $\varphi_0(x) = y_0$ for all $x \in I$. Obviously $\varphi_0(I) \subset J$ and the function $t \mapsto f(t, \varphi_0(t)) = f(t, y_0)$ is continuous on I , so we can define $\varphi_1 : I \rightarrow \mathbb{R}$ by the formula

$$\varphi_1(x) = y_0 + \int_{x_0}^x f(t, \varphi_0(t)) dt \quad (x \in I),$$

which is a continuously differentiable function satisfying

$$\varphi_1(x_0) = y_0 \quad \text{and} \quad \varphi_1'(t) = f(t, \varphi_0(t)) \quad \text{for all } t \in I.$$

Before defining φ_2 , we check that $\varphi_1(I) \subset J$; indeed, for all $x \in I$,

$$|\varphi_1(x) - y_0| = \left| \int_{x_0}^x f(t, \varphi_0(t)) dt \right| \leq M|x - x_0| \leq Md \leq s$$

(the last inequality by the definition of d), whence $\varphi_1(x) \in [y_0 - s, y_0 + s] = J$.

Since $\varphi_1(I) \subset J$ and the function $t \mapsto f(t, \varphi_1(t))$ is continuous on I , we can define $\varphi_2 : I \rightarrow \mathbb{R}$ by the formula

$$\varphi_2(x) = y_0 + \int_{x_0}^x f(t, \varphi_1(t)) dt \quad (x \in I),$$

which is a continuously differentiable function satisfying

$$\varphi_2(x_0) = y_0 \quad \text{and} \quad \varphi_2'(t) = f(t, \varphi_1(t)) \quad \text{for all } t \in I.$$

Suppose φ_n already defined by this procedure,

$$(1) \quad \varphi_n(x) = y_0 + \int_{x_0}^x f(t, \varphi_{n-1}(t)) dt \quad (x \in I),$$

so that

$$(2) \quad \varphi_n(x_0) = y_0 \quad \text{and} \quad \varphi_n'(t) = f(t, \varphi_{n-1}(t)) \quad \text{for all } t \in I.$$

Then $\varphi_n(I) \subset J$ by the computation

$$(3) \quad |\varphi_n(x) - y_0| = \left| \int_{x_0}^x f(t, \varphi_{n-1}(t)) dt \right| \leq M|x - x_0| \leq Md \leq s,$$

and one can define

$$\varphi_{n+1}(x) = y_0 + \int_{x_0}^x f(t, \varphi_n(t)) dt \quad (x \in I),$$

which completes the recursive construction of a sequence (φ_n) satisfying (1) and hence (2).

The proof that (φ_n) is uniformly convergent will follow from the inequalities

$$(4) \quad |\varphi_n(x) - \varphi_{n-1}(x)| \leq \frac{MK^{n-1}}{n!} |x - x_0|^n \quad (x \in I, \ n = 1, 2, 3, \dots),$$

where K is the Lipschitz constant of the assumption (*) in the statement of the theorem. The proof of (4) is by induction on n . For $n = 1$ the inequality (4) follows from the earlier computation

$$|\varphi_1(x) - \varphi_0(x)| = |\varphi_1(x) - y_0| \leq M|x - x_0| = \frac{MK^0}{1!} |x - x_0|^1.$$

Assume inductively that (4) holds for n . Let $x \in I$. Then

$$|\varphi_{n+1}(x) - \varphi_n(x)| = \left| \int_{x_0}^x [f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))] dt \right|,$$

where, by the Lipschitz condition and the induction hypothesis,

$$\begin{aligned} |f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))| &\leq K|\varphi_n(t) - \varphi_{n-1}(t)| \\ &\leq K \cdot \frac{MK^{n-1}}{n!} |t - x_0|^n \\ &= \frac{MK^n}{n!} |t - x_0|^n \end{aligned}$$

for all $t \in I$.

case 1: $x \geq x_0$.

Then, for all $t \in [x_0, x]$, we have $|t - x_0|^n = (t - x_0)^n$ and integration of the preceding displayed inequality yields

$$\begin{aligned} \int_{x_0}^x |f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))| dt &\leq \frac{MK^n}{n!} \int_{x_0}^x (t - x_0)^n dt \\ &= \frac{MK^n}{n!} \cdot \frac{(t - x_0)^{n+1}}{n+1} \Big|_{x_0}^x \\ &= \frac{MK^n}{(n+1)!} (x - x_0)^{n+1} \\ &= \frac{MK^n}{(n+1)!} |x - x_0|^{n+1}, \end{aligned}$$

whence

$$|\varphi_{n+1}(x) - \varphi_n(x)| \leq \int_{x_0}^x |f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))| dt \leq \frac{MK^n}{(n+1)!} |x - x_0|^{n+1}$$

as desired.

case 2: $x \leq x_0$.

By convention $\int_{x_0}^x = -\int_x^{x_0}$, therefore

$$\begin{aligned} |\varphi_{n+1}(x) - \varphi_n(x)| &= \left| \int_{x_0}^x [f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))] dt \right| \\ &\leq \int_x^{x_0} |f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))| dt; \end{aligned}$$

the earlier inequality established for the last integrand (just before case 1) yields, for all $t \in [x, x_0]$, the inequality

$$|f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))| \leq \frac{MK^n}{n!} (x_0 - t)^n,$$

whence

$$|\varphi_{n+1}(x) - \varphi_n(x)| \leq \frac{MK^n}{n!} \int_x^{x_0} (x_0 - t)^n dt$$

and the verification of (4) for $n+1$ continues similarly to case 1.

Since $|x - x_0| \leq d$ for all $x \in I$, it follows from (4) that

$$\|\varphi_n - \varphi_{n-1}\|_\infty \leq \frac{MK^{n-1}}{n!} \cdot d^n = \frac{M}{K} \cdot \frac{(Kd)^n}{n!};$$

since the sequence of positive constants on the right is summable (with sum equal to $\frac{M}{K}[e^{Kd} - 1]$), it follows from the Weierstrass test (6.2.8) that the sequence

$$\sum_{k=1}^n [\varphi_k - \varphi_{k-1}] = \varphi_n - \varphi_0$$

is uniformly convergent, therefore so is the sequence (φ_n) . Let

$$\varphi = \lim \varphi_n$$

be the uniform limit of the sequence, that is, the limit calculated in the space $\mathcal{C}_{\mathbb{R}}(I)$ equipped with the sup-norm metric (6.2.23). Since $\varphi_n(x_0) = y_0$ and $\varphi_n(I) \subset J$ for all n , it follows that $\varphi(x_0) = y_0$ and $\varphi(I) \subset J$.

We wish to pass to the limit in (1) to obtain the analogous formula for φ :

$$(1') \quad \varphi(x) = y_0 + \int_{x_0}^x f(t, \varphi(t)) dt \quad (x \in I).$$

At any rate, it follows from $\varphi(I) \subset J$ and the continuity of f and φ that the function $x \mapsto f(x, \varphi(x))$ ($x \in I$) is defined and continuous on I . Let us write

$$\begin{aligned} g_n(x) &= f(x, \varphi_n(x)) \quad (x \in I, n = 1, 2, 3, \dots) \\ g(x) &= f(x, \varphi(x)) \quad (x \in I). \end{aligned}$$

Then $g_n, g \in \mathcal{C}_{\mathbb{R}}(I)$ and

$$|g_n(x) - g(x)| \leq K|\varphi_n(x) - \varphi(x)| \quad \text{for all } x \in I,$$

by the Lipschitz condition (*), therefore

$$\|g_n - g\|_\infty \leq K\|\varphi_n - \varphi\|_\infty \rightarrow 0,$$

thus $g_n \rightarrow g$ uniformly on I . If $x \in I$ then

$$\begin{aligned} x \geq x_0 &\Rightarrow g_n \rightarrow g \text{ uniformly on } [x_0, x] \\ x \leq x_0 &\Rightarrow g_n \rightarrow g \text{ uniformly on } [x, x_0] \end{aligned}$$

and in either case we conclude that

$$\int_{x_0}^x g_n(t) dt \rightarrow \int_{x_0}^x g(t) dt;$$

for example, if $x \geq x_0$ then

$$\begin{aligned} \left| \int_{x_0}^x g_n(t) dt - \int_{x_0}^x g(t) dt \right| &= \left| \int_{x_0}^x [g_n(t) - g(t)] dt \right| \\ &\leq \int_{x_0}^x |g_n(t) - g(t)| dt \leq (x - x_0) \|g_n - g\|_\infty \rightarrow 0. \end{aligned}$$

Thus, passage to the limit in (1) yields the desired formula (1'), from which it follows that φ is continuously differentiable and $\varphi'(t) = f(t, \varphi(t))$ for all $t \in I$. This completes the proof of the *existence* of a function $\varphi : I \rightarrow \mathbb{R}$ satisfying (**).

Uniqueness: Suppose $\psi : I \rightarrow \mathbb{R}$ is another continuously differentiable function such that $\psi(x_0) = y_0$, $\psi(I) \subset J$ and $\psi'(x) = f(x, \psi(x))$ for all $x \in I$; we are to show that $\psi = \varphi$. By the fundamental theorem of calculus,

$$(5) \quad \psi(x) = y_0 + \int_{x_0}^x f(t, \psi(t)) dt \quad \text{for all } x \in I.$$

Let

$$N = \|\psi - \varphi\|_\infty = \sup_{x \in I} |\psi(x) - \varphi(x)|;$$

at the end of the argument we will know that $N = 0$, but for the moment it suffices that N is a finite upper bound for the values of $\psi - \varphi$. From (5) and the analogous formula for φ , we have

$$\begin{aligned} (6) \quad |\psi(x) - \varphi(x)| &= \left| \int_{x_0}^x [f(t, \psi(t)) - f(t, \varphi(t))] dt \right| \\ &\leq \left| \int_{x_0}^x |f(t, \psi(t)) - f(t, \varphi(t))| dt \right| \end{aligned}$$

for all $x \in I$ (the absolute values around the last integral allow for the possibility that $x < x_0$). By the Lipschitz condition (*),

$$(7) \quad |f(t, \psi(t)) - f(t, \varphi(t))| \leq K|\psi(t) - \varphi(t)| \leq KN,$$

for all $t \in I$; integrating (7) over the interval with endpoints x_0 and x , it follows that

$$\left| \int_{x_0}^x |f(t, \psi(t)) - f(t, \varphi(t))| dt \right| \leq KN|x - x_0| \quad \text{for all } x \in I,$$

therefore, by (6),

$$(8) \quad |\psi(x) - \varphi(x)| \leq KN|x - x_0| \quad \text{for all } x \in I.$$

We assert that for every positive integer n ,

$$(9) \quad |\psi(x) - \varphi(x)| \leq \frac{K^n N |x - x_0|^n}{n!} \quad \text{for all } x \in I.$$

The proof of (9) is by induction on n . For $n = 1$, this is just (8). Assume inductively that (9) holds for n . Fix $x \in I$. Suppose, for example, that $x \geq x_0$. By assumption,

$$|\psi(t) - \varphi(t)| \leq \frac{K^n N (t - x_0)^n}{n!} \quad \text{for all } t \in [x_0, x];$$

substituting this into (7), we have

$$|f(t, \psi(t)) - f(t, \varphi(t))| \leq K \cdot \frac{K^n N (t - x_0)^n}{n!} = \frac{K^{n+1} N (t - x_0)^n}{n!}$$

for all $t \in [x_0, x]$, therefore, by (6),

$$\begin{aligned} |\psi(x) - \varphi(x)| &\leq \frac{K^{n+1} N}{n!} \int_{x_0}^x (t - x_0)^n dt \\ &= \frac{K^{n+1} N}{n!} \cdot \frac{(x - x_0)^{n+1}}{n+1} = \frac{K^{n+1} N (x - x_0)^{n+1}}{(n+1)!}, \end{aligned}$$

thus (9) holds for $n+1$. (When $x < x_0$, the argument is modified in the obvious way, via the convention $\int_{x_0}^x = -\int_x^{x_0}$.) Since $|x - x_0| \leq d$ for all $x \in I$, we see from (9) that

$$|\psi(x) - \varphi(x)| \leq \frac{K^n N d^n}{n!} \quad \text{for all } x \in I,$$

whence

$$N = \sup_{x \in I} |\psi(x) - \varphi(x)| \leq N \cdot \frac{(Kd)^n}{n!};$$

since the right member of the inequality tends to 0 as $n \rightarrow \infty$, we conclude that $N = 0$, that is, $\psi = \varphi$. \diamond

Exercises

1. Assume $f : A \rightarrow \mathbb{R}$ satisfies the hypotheses of Theorem 8.2.1. Then, for each $x \in [x_0 - r, x_0 + r]$, the partial derivative $\partial f / \partial y$ exists at (x, y) for almost every $y \in [y_0 - s, y_0 + s]$.

{Hint: Remark 5.1.10, (vi) and Corollary 5.9.4.}

2. Let $A = [x_0 - r, x_0 + r] \times [y_0 - s, y_0 + s]$ as in Theorem 8.2.1 and let $f : A \rightarrow \mathbb{R}$ be a continuous function. If (i) $\partial f / \partial y$ exists on A , that is,

if for each $x \in [x_0 - r, x_0 + r]$ the function $y \mapsto f(x, y)$ is differentiable on $[y_0 - s, y_0 + s]$ (one-sided at the endpoints), and if (ii) $\partial f / \partial y$ is continuous (or merely bounded) on A , then all hypotheses of Theorem 8.2.1 are fulfilled.

{Hint: Mean-value theorem.}

8.3. Peano's Existence Theorem for $y' = f(x, y)$

The statement of Peano's theorem is obtained from that of Picard's theorem (8.2.1) by the following modifications: the Lipschitz condition (*) is omitted from the hypothesis, and uniqueness is omitted from the conclusion:

8.3.1. Theorem. (G. Peano)¹ *Let $f : A \rightarrow \mathbb{R}$ be a continuous real-valued function defined on a compact rectangle*

$$A = [x_0 - r, x_0 + r] \times [y_0 - s, y_0 + s]$$

in the Euclidean plane \mathbb{R}^2 . Let $M > 0$ be a bound for f on A ,

$$|f(x, y)| \leq M \quad \text{for all } (x, y) \in A,$$

let $d = \min\{r, s/M\}$, and write

$$I = [x_0 - d, x_0 + d], \quad J = [y_0 - s, y_0 + s].$$

Then, there exists a continuously differentiable function $\varphi : I \rightarrow \mathbb{R}$ such that

$$(**) \quad \varphi(x_0) = y_0, \quad \varphi(I) \subset J \quad \text{and} \quad \varphi'(x) = f(x, \varphi(x)) \quad \text{for all } x \in I$$

(where, at the endpoints of I , φ' denotes the one-sided derivative).

The strategy of the proof: (i) approximate f uniformly by a sequence (f_n) of polynomial functions in two real variables x, y (possible by the Stone–Weierstrass theorem); (ii) apply Picard's theorem to each f_n to obtain a solution φ_n of the equation $y' = f_n(x, y)$ with $\varphi_n(x_0) = y_0$; (iii) obtain a solution φ of $y' = f(x, y)$ as the limit of a suitable subsequence of (φ_n) (via Ascoli's theorem). The following lemma assures the Lipschitz condition needed for carrying out step (ii):

¹ Giuseppe Peano (1858–1932). For an extension to vector-valued functions f defined on subsets of a Euclidean space, see the book of Birkhoff and Rota [*op. cit.*, p. 166, Theorem 13]; questions of uniqueness are discussed in the exercises there. Explicit examples of equations with non-unique solutions are given in Jane Cronin's *Differential equations: Introduction and qualitative theory* [2nd edn., Marcel Dekker, New York, 1994], pp. 33–34, and in Philip Hartman's *Ordinary differential equations* [2nd edn., Birkhäuser, Boston, 1982], pp. 18–23.

8.3.2. Lemma. Let S be a nonempty subset of the Cartesian plane \mathbb{R}^2 and let \mathcal{L} be the set of all real-valued functions $f : S \rightarrow \mathbb{R}$ that satisfy a Lipschitz condition with respect to the second variable, that is, for which there is a constant $K_f \geq 0$ such that

$$(*) \quad (x, y_1), (x, y_2) \in S \Rightarrow |f(x, y_1) - f(x, y_2)| \leq K_f |y_1 - y_2|.$$

Then:

(i) \mathcal{L} is a real vector space (for the pointwise linear operations) that contains the constant functions.

(ii) The set $\mathcal{L} \cap \mathcal{B}_{\mathbb{R}}(S)$ of all bounded functions in \mathcal{L} is an algebra (for the pointwise product and linear operations) containing the constant functions.

(iii) If S is a bounded subset of \mathbb{R}^2 then the algebra $\mathcal{L} \cap \mathcal{B}_{\mathbb{R}}(S)$ contains (the restriction to S of) every polynomial function of two real variables x, y .

Proof. (i) Of course \mathcal{L} contains the constant functions on S ; in particular, $0 \in \mathcal{L}$. Let $f, g \in \mathcal{L}$ and $c \in \mathbb{R}$; we are to show that $f + g$ and cf belong to \mathcal{L} . Let K_f and K_g be constants ≥ 0 satisfying $(*)$ for f and g , respectively. It follows from the identities

$$\begin{aligned} (f + g)(x, y_1) - (f + g)(x, y_2) &= [f(x, y_1) - f(x, y_2)] + [g(x, y_1) - g(x, y_2)] \\ (cf)(x, y_1) - (cf)(x, y_2) &= c[f(x, y_1) - f(x, y_2)] \end{aligned}$$

that $f + g$ and cf satisfy $(*)$ with the constants $K_{f+g} = K_f + K_g$ and $K_{cf} = |c|K_f$.

(ii) We know from Example 3.1.10, (i) that $\mathcal{B}_{\mathbb{R}}(S)$ is a real vector space for the pointwise linear operations; moreover, if $f, g \in \mathcal{B}_{\mathbb{R}}(S)$ then the pointwise product fg is also bounded, with $\|fg\|_{\infty} \leq \|f\|_{\infty}\|g\|_{\infty}$ (by the same argument as in Theorem 6.8.12), thus $\mathcal{B}_{\mathbb{R}}(S)$ is an algebra for the pointwise operations. On the other hand, \mathcal{L} is a real vector space by (i), and both \mathcal{L} and $\mathcal{B}_{\mathbb{R}}(S)$ contain the constant functions. Thus we need only show that if $f, g \in \mathcal{L} \cap \mathcal{B}_{\mathbb{R}}(S)$ then the pointwise product fg belongs to \mathcal{L} ; indeed, the computation

$$\begin{aligned} |(fg)(x, y_1) - (fg)(x, y_2)| &= |f(x, y_1)[g(x, y_1) - g(x, y_2)] \\ &\quad + [f(x, y_1) - f(x, y_2)]g(x, y_2)| \\ &\leq \|f\|_{\infty}K_g|y_1 - y_2| + K_f|y_1 - y_2|\|g\|_{\infty} \end{aligned}$$

shows that fg satisfies $(*)$ with $K_{fg} = \|f\|_{\infty}K_g + K_f\|g\|_{\infty}$.

(iii) The hypothesis on S is that there exists a constant $M > 0$ such that $\|(x, y)\|_2 \leq M$ for all $(x, y) \in S$, where $\|(x, y)\|_2 = (x^2 + y^2)^{1/2}$ is the Euclidean norm on \mathbb{R}^2 (3.1.15). Let $c \in \mathbb{R}$ and let u and v be the real-valued functions on S defined by the formulas

$$u(x, y) = x, \quad v(x, y) = y$$

(i.e., the coordinate projection functions of \mathbb{R}^2 restricted to S). Since $|x| \leq \|(x, y)\|_2$ and $|y| \leq \|(x, y)\|_2$, we see that u and v are bounded functions, and the computations

$$\begin{aligned} |u(x, y_1) - u(x, y_2)| &= |x - x| = 0 = 0 \cdot |y_1 - y_2| \\ |v(x, y_1) - v(x, y_2)| &= 1 \cdot |y_1 - y_2| \end{aligned}$$

show that $u, v \in \mathcal{L}$ (with $K_u = 0$ and $K_v = 1$), thus $u, v \in \mathcal{L} \cap \mathcal{B}_{\mathbb{R}}(S)$. In view of (ii), $\mathcal{L} \cap \mathcal{B}_{\mathbb{R}}(S)$ is an algebra containing u, v and the constant functions; the subalgebra of $\mathcal{L} \cap \mathcal{B}_{\mathbb{R}}(S)$ generated by u, v and the constant functions is precisely the set of all functions on S that are polynomials in x and y with real coefficients. \diamond

Proof of Theorem 8.3.1: Brushing past the trivial case that f is identically 0, we can suppose that $M = \|f\|_{\infty} > 0$.

Let $\mathcal{C} = \mathcal{C}_{\mathbb{R}}(A)$ be the real Banach algebra of continuous real-valued functions on the compact rectangle A (6.8.14), let $u, v \in \mathcal{C}$ be the functions defined by $u(x, y) = x$ and $v(x, y) = y$, and let \mathcal{A} be the subalgebra of \mathcal{C} generated by u, v and the constant function 1; as noted in the preceding lemma, \mathcal{A} is the set of all functions on A that are polynomials in x and y ($(x, y) \in A$) with real coefficients. The algebra \mathcal{A} separates the points of A ; indeed, if $(x_1, y_1) \neq (x_2, y_2)$ then either $x_1 \neq x_2$ or $y_1 \neq y_2$, in other words, $u(x_1, y_1) \neq u(x_2, y_2)$ or $v(x_1, y_1) \neq v(x_2, y_2)$. It follows from the Stone-Weierstrass theorem (6.9.9) that \mathcal{A} is uniformly dense in \mathcal{C} .

Choose a sequence (f_n) in \mathcal{A} such that $f_n \rightarrow f$ uniformly on A , that is, $\|f_n - f\|_{\infty} \rightarrow 0$. Then $\|f_n\|_{\infty} \rightarrow \|f\|_{\infty} > 0$; for, if d_{∞} denotes the metric $d_{\infty}(g, h) = \|g - h\|_{\infty}$ derived from the sup-norm, then

$$\|f_n\|_{\infty} = d_{\infty}(f_n, 0) \rightarrow d_{\infty}(f, 0) = \|f\|_{\infty}$$

by Corollary 3.2.4. Suppressing at most finitely many terms, we can suppose that $\|f_n\|_{\infty} > 0$ for all n . Writing $c_n = \|f\|_{\infty} / \|f_n\|_{\infty}$, we have $c_n \rightarrow 1$, therefore

$$\begin{aligned} \|c_n f_n - f\|_{\infty} &= \|(c_n - 1)f_n + (f_n - f)\|_{\infty} \\ &\leq |c_n - 1| \|f_n\|_{\infty} + \|f_n - f\|_{\infty} \\ &\rightarrow 0 \cdot \|f\|_{\infty} + 0 = 0, \end{aligned}$$

that is, $c_n f_n \rightarrow f$ uniformly; moreover,

$$\|c_n f_n\|_{\infty} = c_n \|f_n\|_{\infty} = \|f\|_{\infty} \quad \text{for all } n.$$

Thus, replacing f_n by $c_n f_n$, we can suppose that

$$(1) \quad \|f_n\|_{\infty} = \|f\|_{\infty} = M \quad \text{for all } n.$$

By the Lemma, every f_n satisfies a Lipschitz condition in the second variable y , hence satisfies the hypotheses of Picard's theorem (8.2.1). The significance of the normalization (1) is that the interval I defined for f

in the present theorem is the same as the interval I defined in Picard's theorem for every f_n . Thus, for each n , Picard's theorem provides a continuously differentiable function $\varphi_n : I \rightarrow \mathbb{R}$ such that

$$\varphi_n(x_0) = y_0, \quad \varphi_n(I) \subset J \quad \text{and} \quad \varphi'_n(x) = f_n(x, \varphi_n(x)) \quad \text{for all } x \in I$$

(with one-sided derivatives at the endpoints of I).

Preparatory to obtaining a uniformly convergent subsequence of (φ_n) (via Ascoli's theorem) we show that the sequence is equicontinuous (even equi-uniformly continuous—but see 8.1.10). Let $\epsilon > 0$; we seek a $\delta > 0$ such that

$$(2) \quad x_1, x_2 \in I, \quad |x_1 - x_2| < \delta \quad \Rightarrow \quad |\varphi_n(x_1) - \varphi_n(x_2)| < \epsilon \quad \text{for all } n.$$

In considering pairs x_1, x_2 in I , it will suffice to suppose that $x_1 < x_2$. When $x_1 < x_2$, by the mean-value theorem there exists, for each n , a point $t_n \in (x_1, x_2)$ such that

$$\varphi_n(x_1) - \varphi_n(x_2) = \varphi'_n(t_n) \cdot (x_1 - x_2) = f_n(t_n, \varphi_n(t_n)) \cdot (x_1 - x_2),$$

whence $|\varphi_n(x_1) - \varphi_n(x_2)| \leq M|x_1 - x_2|$. If $\delta > 0$ is chosen so that $M\delta < \epsilon$, then δ meets the requirements of (2).

Since $\varphi_n(I) \subset J$ for all n , and since J is compact hence totally bounded (6.1.26), it is trivial that the sequence (φ_n) is pointwise totally bounded (in the sense of Definition 8.1.1). By Ascoli's theorem (8.1.8), (φ_n) has a subsequence (φ_{n_k}) that is uniformly Cauchy.

Passing to the subsequence and changing notations, we can suppose that (φ_n) is uniformly Cauchy in $\mathcal{C}_{\mathbb{R}}(I) = \mathcal{C}(I, \mathbb{R})$. Since \mathbb{R} is complete, (φ_n) converges uniformly to a function $\varphi \in \mathcal{C}_{\mathbb{R}}(I)$ (6.2.22). It follows from the properties of the φ_n that $\varphi(x_0) = y_0$ and $\varphi(I) \subset J$. It remains only to show that φ is continuously differentiable on I and that $\varphi'(x) = f(x, \varphi(x))$ for all $x \in I$.

We know that for each n , φ_n is continuously differentiable on I , $\varphi_n(x_0) = y_0$ and $\varphi'_n(x) = f_n(x, \varphi_n(x))$ for all $x \in I$, thus, by the fundamental theorem of calculus,

$$(3) \quad \varphi_n(x) = y_0 + \int_{x_0}^x f_n(t, \varphi_n(t)) dt \quad \text{for all } x \in I.$$

Our problem is to “pass to the limit under the integral sign”, so as to obtain in the limit

$$(4) \quad \varphi(x) = y_0 + \int_{x_0}^x f(t, \varphi(t)) dt \quad \text{for all } x \in I;$$

it will then follow from (4) that φ is continuously differentiable on I and that it has the properties listed in (**) of the statement of the theorem.

Let $g : I \rightarrow \mathbb{R}$ and $g_n : I \rightarrow \mathbb{R}$ ($n = 1, 2, 3, \dots$) be the functions defined by the formulas

$$g(x) = f(x, \varphi(x)), \quad g_n(x) = f_n(x, \varphi_n(x)) \quad (x \in I).$$

Then g and the g_n belong to $\mathcal{C}_{\mathbb{R}}(I)$ (by the continuity of the functions that figure in their definition); to deduce (4) from (3), we need only show that

$$\int_{x_0}^x g_n(t) dt \rightarrow \int_{x_0}^x g(t) dt \quad \text{for all } x \in I.$$

We know that $\|f_n - f\|_{\infty} \rightarrow 0$ and $\|\varphi_n - \varphi\|_{\infty} \rightarrow 0$; it will suffice to infer that $\|g_n - g\|_{\infty} \rightarrow 0$, for then it will follow that

$$\left| \int_{x_0}^x g_n(t) dt - \int_{x_0}^x g(t) dt \right| \leq \|g_n - g\|_{\infty} |x - x_0| \rightarrow 0$$

for each $x \in I$. Now comes a very clever move: for every pair of indices m, n , consider the identity

$$\begin{aligned} g_n(x) - g(x) &= f_n(x, \varphi_n(x)) - f(x, \varphi(x)) \\ &= f_n(x, \varphi_n(x)) - f_m(x, \varphi_n(x)) \\ &\quad + f_m(x, \varphi_n(x)) - f_m(x, \varphi(x)) \\ &\quad + f_m(x, \varphi(x)) - f(x, \varphi(x)). \end{aligned}$$

We are interested in the left-most member, but we have introduced a parameter m so that in each of the three differences making up the telescoping sum on the right side, just one index changes: the index on f , the index on φ , and the index on f , respectively. It then follows from the triangle inequality that

$$(5) \quad |g_n(x) - g(x)| \leq \|f_n - f_m\|_{\infty} + |f_m(x, \varphi_n(x)) - f_m(x, \varphi(x))| + \|f_m - f\|_{\infty}$$

for all $x \in I$. Given any $\epsilon > 0$, it will suffice to show that $\|g_n - g\|_{\infty} \leq 4\epsilon$ ultimately. Choose an index N such that

$$n \geq N \quad \Rightarrow \quad \|f_n - f\|_{\infty} \leq \epsilon;$$

then (by the triangle inequality)

$$m, n \geq N \quad \Rightarrow \quad \|f_m - f_n\|_{\infty} \leq 2\epsilon$$

and it follows from (5) that

$$(6) \quad m, n \geq N \quad \Rightarrow \quad |g_n(x) - g(x)| \leq 3\epsilon + |f_m(x, \varphi_n(x)) - f_m(x, \varphi(x))|$$

for all $x \in I$. Setting $m = N$ in (6), we have

$$(7) \quad n \geq N \quad \Rightarrow \quad |g_n(x) - g(x)| \leq 3\epsilon + |f_N(x, \varphi_n(x)) - f_N(x, \varphi(x))|;$$

thus, if $K_N > 0$ is a constant such that

$$|f_N(x, y_1) - f_N(x, y_2)| \leq K_N |y_1 - y_2|$$

for all x, y_1, y_2 , it follows from (7) that

$$n \geq N \quad \Rightarrow \quad |g_n(x) - g(x)| \leq 3\epsilon + K_N |\varphi_n(x) - \varphi(x)|$$

for all $x \in I$, whence

$$(8) \quad \|g_n - g\|_\infty \leq 3\epsilon + K_N \|\varphi_n - \varphi\|_\infty \quad \text{for all } n \geq N.$$

Since $\|\varphi_n - \varphi\|_\infty \leq \epsilon/K_N$ ultimately, it follows from (8) that $\|g_n - g\|_\infty \leq 4\epsilon$ ultimately; we have shown that $\|g_n - g\|_\infty \rightarrow 0$, which completes the verification of (4) and hence of (**). \diamond

To appreciate the cunning of the three-term telescoping sum (with a parameter m) in the foregoing proof, contemplate trying to reach the same goal with the following two-term telescoping sum (without the parameter m):

$$\begin{aligned} g_n(x) - g(x) &= f_n(x, \varphi_n(x)) - f(x, \varphi(x)) \\ &= f_n(x, \varphi_n(x)) - f(x, \varphi_n(x)) \\ &\quad + f(x, \varphi_n(x)) - f(x, \varphi(x)); \end{aligned}$$

in the last difference on the right, we have no Lipschitz condition on f to push the proof along. Impasse.

Try again: if K_n is a Lipschitz constant for f_n then, from the formula

$$\begin{aligned} g_n(x) - g(x) &= f_n(x, \varphi_n(x)) - f(x, \varphi(x)) \\ &= f_n(x, \varphi_n(x)) - f_n(x, \varphi(x)) \\ &\quad + f_n(x, \varphi(x)) - f(x, \varphi(x)), \end{aligned}$$

we infer that

$$\begin{aligned} |g_n(x) - g(x)| &\leq K_n |\varphi_n(x) - \varphi(x)| + |f_n(x, \varphi(x)) - f(x, \varphi(x))| \\ &\leq K_n \|\varphi_n - \varphi\|_\infty + \|f_n - f\|_\infty; \end{aligned}$$

we are now blocked by the possibility that the sequence (K_n) may be unbounded.

Thank you, parameter m .

CHAPTER 9

Topics in Measure and Integration

- §9.1. Jordan–Hahn decomposition of a signed measure
- §9.2. Radon–Nikodym theorem
- §9.3. Lebesgue decomposition of measures
- §9.4. Convolution in $L^1(\mathbb{R})$
- §9.5. Integral operators (with continuous kernel function)

In Section 1, the decomposition of a finite signed measure as a difference of finite measures proved in Chapter 4 (4.8.8) is generalized to countably additive set functions admitting either $+\infty$ or $-\infty$ (but not both) as values.

The decomposition theorem of Section 1 is applied in Section 2 to generalize the Radon–Nikodym theorem proved in Chapter 4 for finite measures (4.8.11) to the σ -finite case.

The Radon–Nikodym theorem of Section 2 is applied in Section 3 to show that if ν and μ are any two σ -finite measures on a measurable space (X, \mathcal{S}) , then the underlying set X can be partitioned into three pairwise disjoint measurable sets, $X = E \cup F \cup G$, such that $\mu(F) = \nu(G) = 0$ and such that a measurable subset of E is negligible for μ if and only if it is negligible for ν .

Section 4 is an application of the Fubini–Tonelli theorem of §7.4 (for the product of σ -finite measures): the convolution $f * g$ of two Lebesgue-integrable functions f and g on \mathbb{R} is defined and is shown to have the properties desired of it (for application in the theory of Fourier transforms).

Section 5 is an application of Ascoli's theorem (§8.1): it is shown that if $I = [a, b]$ is a closed interval of \mathbb{R} , then every continuous complex-valued function defined on the square $I \times I$ induces a continuous linear mapping in the Hilbert space $L^2_C(I)$ with respect to Lebesgue measure on I .

9.1. Jordan–Hahn Decomposition of a Signed Measure

The difference $\alpha - \beta$ of two extended real numbers is defined if and only if α and β are not both equal to $+\infty$ nor both equal to $-\infty$ (1.15.4, (v));

let us say in this case that α and β are *subtractible* (a relation that is symmetric in α and β). The concept carries over to extended-real-valued functions defined on a nonempty set T :

9.1.1. *Definition.* Functions $f_1, f_2 : T \rightarrow \overline{\mathbb{R}}$ are said to be **subtractible** if, for every $t \in T$, $f_1(t)$ and $f_2(t)$ are subtractible; the difference function $f_1 - f_2$ is then defined by the formula $(f_1 - f_2)(t) = f_1(t) - f_2(t)$ ($t \in T$).

For the rest of the section, we fix a measurable space (X, \mathcal{S}) , that is, a set X and a σ -algebra \mathcal{S} of subsets of X (4.1.1).

A measure on \mathcal{S} is a nonnegative, extended-real-valued function that is countably additive and vanishes at the empty set (2.4.12); the goal of the present section is to characterize the functions on \mathcal{S} that are differences $\mu_1 - \mu_2$ of pairs of measures on \mathcal{S} . We note first that not all pairs of measures are subtractible:

9.1.2. *Remark.* If μ_1 and μ_2 are measures on \mathcal{S} , then

μ_1, μ_2 are subtractible \Leftrightarrow at least one of μ_1, μ_2 is a *finite* measure.

{Proof: \Leftarrow : Obvious from Definition 9.1.1.

\Rightarrow : Arguing contrapositively, if neither μ_1 nor μ_2 is finite, then $\mu_1(X) = \mu_2(X) = +\infty$, consequently μ_1, μ_2 are not subtractible.}

9.1.3. *Theorem.* If μ_1 and μ_2 are measures on \mathcal{S} , at least one of which is finite, then the function $\nu = \mu_1 - \mu_2$ has the following properties:

1° $\nu(\emptyset) = 0$.

2° If (E_n) is a sequence of pairwise disjoint sets in \mathcal{S} , and if $E = \bigcup_{n=1}^{\infty} E_n$, then

$$\nu(E) = \sum_{n=1}^{\infty} \nu(E_n),$$

in the sense that $\sum_{k=1}^n \nu(E_k)$ is defined for every n and

$$\sum_{k=1}^n \nu(E_k) \rightarrow \nu(E) \quad \text{in } \overline{\mathbb{R}}$$

as $n \rightarrow \infty$.

3° ν does not take on both of the values $+\infty$ and $-\infty$.

Proof. Suppose, for example, that μ_1 is finite. Then ν does not take on the value $+\infty$, whence 3°. Property 1° is obvious. We know from the

countable additivity of μ_1 and μ_2 that

$$\sum_{k=1}^n \mu_1(E_k) \rightarrow \mu_1(E) \quad \text{in } \mathbb{R},$$

$$\sum_{k=1}^n \mu_2(E_k) \rightarrow \mu_2(E) \quad \text{in } \overline{\mathbb{R}}$$

(convergence in $\overline{\mathbb{R}}$ is defined in 1.16.8), so it clearly suffices to prove the following lemma: *If $\alpha_n \rightarrow \alpha$ in \mathbb{R} and $\beta_n \rightarrow \beta$ in $\overline{\mathbb{R}}$, then $\alpha_n + \beta_n \rightarrow \alpha + \beta$ in $\overline{\mathbb{R}}$.* At any rate, all the sums in question are defined, and the sequence (α_n) is bounded. If $\beta \in \mathbb{R}$ then β_n is ultimately bounded (cf. 1.16.6) and, since convergence is undisturbed by amputating a finite number of terms, it is clear that $\alpha_n + \beta_n \rightarrow \alpha + \beta$. If $\beta = +\infty$ then (1.16.9)

$$r \in \mathbb{R}, r > 0 \Rightarrow \text{ultimately } \beta_n > 2r - \alpha \text{ and } \alpha_n > \alpha - r,$$

therefore $\alpha_n + \beta_n > r$ ultimately, thus $\alpha_n + \beta_n \rightarrow +\infty = \alpha + \beta$. Finally, if $\beta = -\infty$ then

$$-\alpha_n \rightarrow -\alpha \text{ in } \mathbb{R} \text{ and } -\beta_n \rightarrow +\infty \text{ in } \overline{\mathbb{R}},$$

therefore $-\alpha_n + (-\beta_n) \rightarrow +\infty$ by the preceding case, whence $\alpha_n + \beta_n \rightarrow -\infty = \alpha + \beta$. \diamond

9.1.4. Definition. A **signed measure** on \mathcal{S} is a function $\nu : \mathcal{S} \rightarrow \overline{\mathbb{R}}$ satisfying the conditions 1° and 2° of the above theorem; the property 2° is expressed by saying that ν is **countably additive**. (The property 3° is automatically verified, as we shall see in 9.1.7 below.)

9.1.5. Example. The *real measures* on \mathcal{S} considered in §6.5 are precisely the signed measures all of whose values are finite; in other words, they are the *finite signed measures* discussed in §4.8.

For the rest of the section, ν denotes a signed measure on \mathcal{S} .

Our main objective is to show that ν can be expressed as a difference of two measures (at least one of them finite) as in Theorem 9.1.3, a result known as the Jordan–Hahn decomposition theorem. We prepare the way for this by establishing some basic properties of ν in a series of six propositions.

9.1.6. Proposition. ν is *finitely additive*, that is, if E_1, \dots, E_r are pairwise disjoint sets in \mathcal{S} , then

$$\nu(E_1 \cup \dots \cup E_r) = \sum_{k=1}^r \nu(E_k)$$

(in particular, the sum on the right side exists in $\overline{\mathbb{R}}$).

Proof. Define $E_k = \emptyset$ for every positive integer $k > r$ and apply property 2° of ν to the sequence (E_n) . \diamond

9.1.7. Proposition. ν satisfies 3° of 9.1.3. In particular, the sum $\nu(E) + \nu(F)$ exists in $\overline{\mathbb{R}}$ for every pair of sets E, F in \mathcal{S} .

*Proof.*¹ Assume to the contrary that $\nu(E) = +\infty$ and $\nu(F) = -\infty$ for suitable sets E, F in \mathcal{S} . Consider the following decompositions of $E \cup F$:

$$E \cup F = E \cup (E' \cap F) = (E \cap F') \cup F.$$

By the finite additivity of ν (9.1.6),

$$\begin{aligned}\nu(E \cup F) &= \nu(E) + \nu(E' \cap F) = \nu(E \cap F') + \nu(F) \\ &= +\infty + \nu(E' \cap F) = \nu(E \cap F') + (-\infty)\end{aligned}$$

(in particular, the indicated sums in $\overline{\mathbb{R}}$ exist), whence the absurdity $\nu(E \cup F) = +\infty = -\infty$. \diamond

9.1.8. Proposition. If $E \in \mathcal{S}$ and $\nu(E) \in \mathbb{R}$, then

$$F \in \mathcal{S}, F \subset E \Rightarrow \nu(F) \in \mathbb{R},$$

hence also

$$F \in \mathcal{S} \Rightarrow \nu(E \cap F) \in \mathbb{R}.$$

Proof. Assuming $F \in \mathcal{S}$ and $F \subset E$, so that $E = (E - F) \cup F$, then

$$\nu(E) = \nu(E - F) + \nu(F)$$

by the additivity of ν ; in particular, the sum is defined in $\overline{\mathbb{R}}$ and is equal to the real number $\nu(E)$, consequently both terms of the sum must be in \mathbb{R} . The second implication is immediate from the first. \diamond

9.1.9. Proposition. Let (E_n) be a sequence of pairwise disjoint sets in \mathcal{S} and let $E = \bigcup_{n=1}^{\infty} E_n$, so that

$$\nu(E) = \sum_{n=1}^{\infty} \nu(E_n)$$

by the condition 2° of 9.1.4. If $\nu(E) \in \mathbb{R}$ then $\nu(E_n) \in \mathbb{R}$ for all n , and the series is absolutely convergent.

Proof. The terms $\nu(E_n)$ of the series are real numbers by 9.1.8, and the convergence is absolute by the same argument as for 4.8.3, (i).² \diamond

¹ E. Hewitt and K. Stromberg, *Real and abstract analysis* [Springer, New York, 1965], p. 304, (19.2).

² As remarked in the proof of 4.8.3, (i), a ‘commutatively convergent’ infinite series of real numbers (i.e., a series which is convergent for every permutation of its terms) is absolutely convergent; for a straightforward elementary proof, see E. Landau, *Differential and integral calculus* [Chelsea, New York, 1951], p. 158, Theorem 217.

9.1.10. Proposition. Let (E_n) be a sequence of sets in \mathcal{S} and let $E \in \mathcal{S}$.

(i) If $E_n \uparrow E$ then $\nu(E_n) \rightarrow \nu(E)$ in \mathbb{R} .

(ii) If $E_n \downarrow E$ and $\nu(E_1) \in \mathbb{R}$, then $\nu(E_n) \rightarrow \nu(E)$ in \mathbb{R} .

Proof. (i) Define $E_0 = \emptyset$ and $F_n = E_n - E_{n-1}$ for every positive integer n . Then (F_n) is a sequence of pairwise disjoint sets in \mathcal{S} with union E , consequently

$$\nu(E) = \sum_{n=1}^{\infty} \nu(F_n);$$

since, for each n ,

$$\sum_{k=1}^n \nu(F_k) = \nu\left(\bigcup_{k=1}^n F_k\right) = \nu(E_n),$$

this means that $\nu(E_n) \rightarrow \nu(E)$ as $n \rightarrow \infty$ (9.1.4).

(ii) By 9.1.8, $\nu(E)$ and the $\nu(E_n)$ are real numbers. In particular, from $E_1 = (E_1 - E) \cup E$ and the additivity of ν , we infer that $\nu(E_1 - E) = \nu(E_1) - \nu(E)$. Writing $G_n = E_1 - E_n$, we have $E_1 = G_n \cup E_n$ and

$$\nu(G_n) + \nu(E_n) = \nu(E_1) \in \mathbb{R},$$

consequently $\nu(G_n) = \nu(E_1) - \nu(E_n)$; since $G_n \uparrow E_1 - E$, it follows from (i) that

$$\nu(G_n) \rightarrow \nu(E_1 - E) \text{ in } \mathbb{R},$$

that is,

$$\nu(E_1) - \nu(E_n) \rightarrow \nu(E_1) - \nu(E),$$

whence $\nu(E_n) \rightarrow \nu(E)$. \diamond

9.1.11. Proposition. Suppose E is a set in \mathcal{S} such that $\nu(E)$ is a real number. If $(E_i)_{i \in I}$ is a pairwise disjoint family of sets in \mathcal{S} such that $E_i \subset E$ for all $i \in I$, then $\nu(E_i) = 0$ for all but countably many indices i , that is, the set

$$J = \{i \in I : \nu(E_i) \neq 0\}$$

is a countable subset of I .

Proof. As noted in 9.1.8, $\nu(E_i) \in \mathbb{R}$ for all $i \in I$. For each positive integer n , let

$$J_n = \{i \in I : |\nu(E_i)| \geq 1/n\};$$

clearly $J_n \uparrow J$, so it will suffice to show that every J_n is finite.

Assume to the contrary that J_n is infinite for some index n . Choose a sequence i_1, i_2, i_3, \dots of distinct indices in J_n (1.9.4) and let

$$F = \bigcup_{k=1}^{\infty} E_{i_k};$$

then $F \in \mathcal{S}$ and $F \subset E$, therefore $\nu(F) \in \mathbb{R}$. Since the E_{i_k} are pairwise disjoint, we infer from 9.1.9 that the series

$$\sum_{k=1}^{\infty} |\nu(E_{i_k})|$$

is convergent, contrary to the fact that $|\nu(E_{i_k})| \geq 1/n$ for all k (because $i_k \in J_n$). \diamond

The following definition extends Definition 4.8.4 (the special case of a finite signed measure):

9.1.12. *Definition.* Let $A \in \mathcal{S}$. Define a set function $\nu_A : \mathcal{S} \rightarrow \overline{\mathbb{R}}$ by the formula

$$\nu_A(E) = \nu(A \cap E) \quad (E \in \mathcal{S}).$$

The basic formal properties of this notation are gathered in the following proposition:

9.1.13. *Proposition.* Let $A, B, C \in \mathcal{S}$.

- (i) $\nu_{\emptyset} = 0$.
- (ii) ν_A is a signed measure on \mathcal{S} .
- (iii) ν_A is a finite signed measure $\Leftrightarrow \nu(A) \in \mathbb{R}$.
- (iv) $(\nu_A)_B = \nu_{A \cap B}$.
- (v) $A \cap B = \emptyset \Rightarrow \nu_{A \cup B} = \nu_A + \nu_B$.
- (vi) If $\nu(A) \in \mathbb{R}$ then $\nu_{X-A} = \nu - \nu_A$.
- (vii) If either $\nu(A) \in \mathbb{R}$ or $\nu(B) \in \mathbb{R}$ then

$$\nu_{A \cup B} + \nu_{A \cap B} = \nu_A + \nu_B.$$

Proof. (i) Obvious.

(ii) The countable additivity of ν_A follows from that of ν and the identity

$$A \cap \left(\bigcup_{k=1}^{\infty} E_k \right) = \bigcup_{k=1}^{\infty} A \cap E_k.$$

(iii) If ν_A is finite then in particular $\nu(A) = \nu(A \cap X) = \nu_A(X) \in \mathbb{R}$; conversely, if $\nu(A) \in \mathbb{R}$ then $\nu(A \cap E) \in \mathbb{R}$ for all $E \in \mathcal{S}$ by 9.1.8, thus ν_A is finite.

(iv) $\nu_{A \cap B}(E) = \nu(A \cap B \cap E) = \nu_A(B \cap E) = (\nu_A)_B(E)$ for all $E \in \mathcal{S}$.

(v) Clear from the additivity of ν .

(vi) By (v), $\nu = \nu_X = \nu_{X-A} + \nu_A$, and the term ν_A is transposable by (iii).

(vii) Suppose $\nu(A) \in \mathbb{R}$, so that ν_A and $\nu_{A \cap B}$ are finite signed measures. From $A \cup B - A = B - A \cap B$, we have

$$(*) \quad \nu_{A \cup B - A} = \nu_{B - A \cap B}.$$

But $A \cup B = (A \cup B - A) \cup A$ yields

$$\nu_{A \cup B} = \nu_{A \cup B - A} + \nu_A,$$

whence $\nu_{A \cup B - A} = \nu_{A \cup B} - \nu_A$, and similarly $B = (B - A \cap B) \cup (A \cap B)$ yields

$$\nu_B = \nu_{B - A \cap B} + \nu_{A \cap B},$$

whence $\nu_{B - A \cap B} = \nu_B - \nu_{A \cap B}$; substituting these equations into (*), we have

$$\nu_{A \cup B} - \nu_A = \nu_B - \nu_{A \cap B},$$

and the finite signed measures ν_A and $\nu_{A \cap B}$ are transposable. \diamond

9.1.14. *Remarks.* Let $A \in \mathcal{S}$. Writing

$$A \cap \mathcal{S} = \{A \cap E : E \in \mathcal{S}\} = \{E \in \mathcal{S} : E \subset A\},$$

it is easy to see that $A \cap \mathcal{S}$ is a σ -algebra of subsets of A and that the restriction $\nu|_{A \cap \mathcal{S}}$ of ν to $A \cap \mathcal{S}$ is a signed measure in the context of the measurable space $(A, A \cap \mathcal{S})$.

9.1.15. *Definition.* With the preceding notations, the signed measure $\nu|_{A \cap \mathcal{S}}$ is abbreviated $\nu|_A$. (Abuse of notation: $\nu|_A$ is not a function on A .)

9.1.16. *Definition.* A set $A \in \mathcal{S}$ is said to be **purely positive** with respect to ν if $\nu_A \geq 0$, that is, if ν_A is a measure on \mathcal{S} (equivalently, $\nu|_A$ is a measure on $A \cap \mathcal{S}$), and we then write $A \geq 0$ (with respect to ν). Similarly, a set $A \in \mathcal{S}$ is said to be **purely negative** with respect to ν , written $A \leq 0$, if $\nu_A \leq 0$ (equivalently, $-\nu_A$ is a measure on \mathcal{S}).

9.1.17. *Remarks.* The preceding notations have the following properties:

- (i) $\nu_A = 0 \Leftrightarrow A \geq 0 \text{ \& } A \leq 0$.
- (ii) $\nu \geq 0 \Rightarrow \nu_A \geq 0$ for all $A \in \mathcal{S}$.
- (iii) $A \geq 0 \Rightarrow A \cap B \geq 0$ for all $B \in \mathcal{S}$.
- (iv) $A \geq 0 \text{ \& } B \geq 0 \Rightarrow A \cup B \geq 0$.
- (v) $A_n \geq 0$ ($n = 1, 2, 3, \dots$) $\Rightarrow \bigcup_{n=1}^{\infty} A_n \geq 0$.

{Proof: (i), (ii) are obvious.

(iii) This follows from (ii) and the formula $\nu_{A \cap B} = (\nu_A)_B$.

(iv) When $A \cap B = \emptyset$ this follows from $\nu_{A \cup B} = \nu_A + \nu_B$, and the general case then follows from (iii) and the formula $A \cup B = A \cup (B \cap A')$.

(v) Let $A = \bigcup_{n=1}^{\infty} A_n$. Replacing A_n by $A_1 \cup \dots \cup A_n$ one can suppose, in view of (iv), that $A_n \uparrow A$; it then follows from 9.1.10 that

$$\nu_A(E) = \nu(A \cap E) = \lim \nu(A_n \cap E) = \lim \nu_{A_n}(E) \geq 0$$

for all $E \in \mathcal{S}$, thus $\nu_A \geq 0$.}

The key “existence theorem” of this section is as follows:

9.1.18. Lemma. *If $A \in \mathcal{S}$ and $0 < \nu(A) < +\infty$, then there exists a set $A_0 \in \mathcal{S}$ such that*

$$A_0 \subset A, \quad A_0 \geq 0 \quad (\text{with respect to } \nu) \quad \text{and} \quad \nu(A_0) > 0 .$$

Proof. If $A \geq 0$, that is, if $\nu_A \geq 0$, then $A_0 = A$ meets the requirements. Otherwise, there exists a set $B \in \mathcal{S}$ with $B \subset A$ and $\nu(B) < 0$. Let $(B_i)_{i \in I}$ be a maximal family of pairwise disjoint sets such that

$$B_i \in \mathcal{S}, \quad B_i \subset A \quad \text{and} \quad \nu(B_i) < 0 \quad \text{for all } i \in I$$

(such a family exists by Zorn’s lemma). By 9.1.11, the index set I is countable, therefore the set $B = \bigcup_{i \in I} B_i$ belongs to \mathcal{S} . Of course $B \subset A$, and

$$\nu(B) = \sum_{i \in I} \nu(B_i) < 0 .$$

Let $A_0 = A - B$. Since $\nu(A) \in \mathbb{R}$, we have $\nu(A_0) = \nu(A) - \nu(B)$ by 9.1.8 and the additivity of ν , and since $\nu(A) > 0$ and $\nu(B) < 0$ we conclude that $\nu(A_0) > 0$.

It remains only to show that $A_0 \geq 0$. Thus, if $E \in \mathcal{S}$ and $E \subset A_0$, we need only show that $\nu(E) \geq 0$. The alternative, $\nu(E) < 0$, would contradict the maximality of the family $(B_i)_{i \in I}$. \diamond

9.1.19. Remark. Application of the lemma to $-\nu$ yields the dual result: If $-\infty < \nu(A) < 0$ then there exists a set $A_0 \in \mathcal{S}$ such that $A_0 \subset A$, $A_0 \leq 0$ (with respect to ν) and $\nu(A_0) < 0$.

All the needed tools are in hand for an efficient proof of the desired decomposition theorem:

9.1.20. Theorem. (Jordan–Hahn decomposition)³ *Let (X, \mathcal{S}) be a measurable space. If ν is any signed measure on the σ -algebra \mathcal{S} , then there exists a set $A \in \mathcal{S}$ such that*

$$A \geq 0 \quad \text{and} \quad X - A \leq 0 \quad (\text{with respect to } \nu) .$$

Defining $\mu_1 = \nu_A$ and $\mu_2 = -\nu_{X-A}$, μ_1 and μ_2 are measures on \mathcal{S} such that $\nu = \mu_1 - \mu_2$.

³ Camille Jordan (1838–1922), Hans Hahn (1879–1934).

Proof. We know that ν does not take on both of the values $+\infty$ and $-\infty$ (9.1.7). Suppose, for example, that the value $+\infty$ is not taken on, so that

$$-\infty \leq \nu(E) < +\infty \quad \text{for all } E \in \mathcal{S}.$$

Let

$$\mathcal{P} = \{A \in \mathcal{S} : A \geq 0 \text{ with respect to } \nu\}$$

(for example, $\emptyset \in \mathcal{P}$). We know from 9.1.17, (v) that \mathcal{P} is closed under countable unions. Moreover, the values of ν on \mathcal{P} are real numbers ≥ 0 .

We assert that ν takes on a largest value on \mathcal{P} . For, let

$$\alpha = \sup\{\nu(B) : B \in \mathcal{P}\}$$

and let (A_n) be a sequence in \mathcal{P} such that $\nu(A_n) \rightarrow \alpha$ in $\overline{\mathbb{R}}$. Replacing A_n by $A_1 \cup \dots \cup A_n$, we can suppose that $A_n \uparrow$. Then, writing $A = \bigcup_{n=1}^{\infty} A_n$, we have $A_n \uparrow A$, therefore $\nu(A_n) \rightarrow \nu(A)$ by 9.1.10, thus

$$\nu(A) = \lim \nu(A_n) = \alpha = \sup\{\nu(B) : B \in \mathcal{P}\};$$

since $A \in \mathcal{P}$, we conclude that ν takes on its largest value at A . In particular, $0 \leq \nu(A) < +\infty$.

Since $A \in \mathcal{P}$, we know that $A \geq 0$. We need only show that $X - A \leq 0$. Assuming to the contrary that there exists a set $E \in \mathcal{S}$ such that $E \subset X - A$ and $\nu(E) > 0$, we then have $0 < \nu(E) < \infty$; by the lemma, there exists a set $A_0 \in \mathcal{S}$ such that $A_0 \subset E$, $A_0 \geq 0$ and $\nu(A_0) > 0$. Thus $A_0 \in \mathcal{P}$ and A_0 is disjoint from A (because $A_0 \subset E \subset X - A$), therefore

$$\nu(A \cup A_0) = \nu(A) + \nu(A_0) > \nu(A) = \alpha;$$

but $A \cup A_0 \in \mathcal{P}$ by 9.1.17, (iv), therefore $\nu(A \cup A_0) \leq \alpha$ by the definition of α , a contradiction. \diamond

9.1.21. *Remark.* The measures μ_1, μ_2 constructed by the method of the preceding theorem are unique. That is, if also $B \in \mathcal{S}$, $B \geq 0$ and $X - B \leq 0$ (with respect to ν) then

$$\nu_A = \nu_B \quad \text{and} \quad \nu_{X-A} = \nu_{X-B}.$$

For, $A \cap B' \geq 0$ (because $A \geq 0$) and $A \cap B' \leq 0$ (because $B' \leq 0$), consequently $\nu_{A \cap B'} = 0$. Similarly $\nu_{A' \cap B} = 0$. From $A = (A \cap B) \cup (A \cap B')$ we infer that

$$\nu_A = \nu_{A \cap B} + \nu_{A \cap B'} = \nu_{A \cap B}$$

and similarly $\nu_B = \nu_{A \cap B}$, thus $\nu_A = \nu_B$. Similarly $\nu_{X-A} = \nu_{X-B}$.

9.1.22. *Definition.* With notations as in Theorem 9.1.20, one writes

$$\nu^+ = \nu_A \quad \text{and} \quad \nu^- = -\nu_{X-A}$$

(the measures ν^+ and ν^- depend only on ν by the preceding remark), and the formula

$$\nu = \nu^+ - \nu^-$$

is called the **Jordan–Hahn decomposition** of the signed measure ν . The measure $\nu^+ + \nu^-$ is called the **total variation** of ν and is denoted

$$|\nu| = \nu^+ + \nu^-.$$

Inspecting the proof of Theorem 9.1.20, we see that if ν does not take on the value $+\infty$ then ν^+ is finite. If ν does not take on the value $-\infty$, then ν^- is finite.

The following proposition will be useful on several occasions in the next two sections:

9.1.23. Proposition. *Let ν be a signed measure on \mathcal{S} . Then:*

- (i) $\nu = 0 \Leftrightarrow |\nu| = 0$.
- (ii) *For every measurable set $E \in \mathcal{S}$,*

$$(\nu_E)^+ = (\nu^+)_E, \quad (\nu_E)^- = (\nu^-)_E, \quad |\nu_E| = |\nu|_E.$$

Proof. (i) If $\nu = 0$ then, with notations as in Definition 9.1.22, $\nu^+ = \nu_A = 0_A = 0$ and similarly $\nu^- = 0$, therefore $|\nu| = \nu^+ + \nu^- = 0$. Conversely, if $|\nu| = 0$ then $\nu^+ = \nu^- = 0$, therefore $\nu = \nu^+ - \nu^- = 0$.

(ii) With notations as in 9.1.22, we have

$$\begin{aligned} (\nu_E)_A &= \nu_{E \cap A} = (\nu_A)_E = (\nu^+)_E \geq 0 \\ (\nu_E)_{X-A} &= \nu_{E \cap (X-A)} = (\nu_{X-A})_E = (-\nu^-)_E \leq 0; \end{aligned}$$

thus, at least one of the measures $(\nu_E)_A$, $-(\nu_E)_{X-A}$ is finite, and A defines a Jordan–Hahn decomposition of the signed measure ν_E , with

$$\begin{aligned} (\nu_E)^+ &= (\nu_E)_A = (\nu^+)_E, \\ (\nu_E)^- &= -(\nu_E)_{X-A} = -(\nu_{X-A})_E = -(-\nu^-)_E = (\nu^-)_E, \\ |\nu_E| &= (\nu_E)^+ + (\nu_E)^- = (\nu^+)_E + (\nu^-)_E = (\nu^+ + \nu^-)_E = |\nu|_E. \quad \diamond \end{aligned}$$

Exercises

1. Let (X, \mathcal{S}) be a measurable space.

(i) The σ -algebra \mathcal{S} is a commutative ring with unity for the operations of sum and product defined, respectively, by the formulas

$$\begin{aligned} E \oplus F &= (E - F) \cup (F - E), \\ E \odot F &= E \cap F, \end{aligned}$$

having \emptyset as zero element, X as unity element, and satisfying $E \odot E = E$ for all $E \in \mathcal{S}$. (The same is true for every algebra of sets.)

(ii) If ν is a signed measure on \mathcal{S} , and if

$$\mathcal{S}_0 = \{E \in \mathcal{S} : \nu(E) \in \mathbb{R}\},$$

then \mathcal{S}_0 is an ideal in \mathcal{S} (for the ring structure just described).

{Hint: 9.1.8.}

(iii) (Theorem of M.H. Stone)⁴ If R is a ring with unity such that $x^2 = x$ for all $x \in R$ (such rings are called *Boolean*) then R is commutative ($xy = yx$ for all x, y in R) and R may be regarded as (i.e., is isomorphic to) an algebra of subsets of a suitable set, with operations as described in (i).

2. Let (X, \mathcal{S}) be a measurable space, ν a signed measure on \mathcal{S} . Define

$$\mathcal{S}_+ = \{A \in \mathcal{S} : \nu_A \geq 0\}$$

$$\mathcal{S}_- = \{A \in \mathcal{S} : \nu_A \leq 0\} = \{A \in \mathcal{S} : (-\nu)_A \geq 0\}$$

$$\mathcal{S}_0 = \{A \in \mathcal{S} : \nu_A = 0\} = \mathcal{S}_+ \cap \mathcal{S}_-.$$

Then \mathcal{S}_+ is a σ -ring of subsets of X (therefore so are \mathcal{S}_- and \mathcal{S}_0), and \mathcal{S}_+ is a σ -algebra if and only if $\nu \geq 0$. {A **ring** of subsets of X is a set \mathcal{R} of subsets of X , with $\emptyset \in \mathcal{R}$, such that if $A, B \in \mathcal{R}$ then also $A - B, A \cup B \in \mathcal{R}$; a ring that is closed under countable unions is called a **σ -ring**.}

9.2. Radon–Nikodym Theorem

Throughout this section, (X, \mathcal{S}, μ) is a measure space. (Later in the section, it will be assumed that μ is σ -finite.)

If $f \in \mathcal{L}^1(\mu)$, that is, if $f : X \rightarrow \mathbb{R}$ is μ -integrable (4.4.7), we know that the indefinite integral $f \cdot \mu : \mathcal{S} \rightarrow \mathbb{R}$, defined by

$$(f \cdot \mu)(E) = \int_E f d\mu \quad (E \in \mathcal{S}),$$

is a real measure on \mathcal{S} (4.7.3) such that

$$E \in \mathcal{S}, \mu(E) = 0 \quad \Rightarrow \quad (f \cdot \mu)(E) = 0$$

(see 4.7.2, (vii)). Thus, for every μ -integrable function $f : X \rightarrow \mathbb{R}$, the set function $\nu = f \cdot \mu$ is a real measure on \mathcal{S} that is absolutely continuous with respect to μ in the sense of Definition 4.8.6 (written $\nu \ll \mu$).

It was shown in Corollary 4.8.12 that if μ is a *finite* measure, then every real measure ν on \mathcal{S} such that $\nu \ll \mu$ has the form $\nu = f \cdot \mu$ for some μ -integrable function f . Our objective in this section is to generalize this result so as to permit μ to be σ -finite and ν to be a signed measure (with

⁴ Cf. P.R. Halmos, *Measure theory* [Van Nostrand, New York, 1950; reprinted Springer-Verlag, New York, 1974], p. 170, Exercise (15a).

possibly infinite values). The passage to σ -finite μ is straightforward, but admitting signed measures poses two technical problems: when ν takes on infinite values, (1) the function f can no longer be required to be μ -integrable—we will need to define $f \cdot \mu$ for certain measurable functions f that are not μ -integrable, and (2) the condition $\nu \ll \mu$ will no longer suffice, but must be augmented with the assumption that the measure $|\nu|$ (defined in 9.1.22) is also σ -finite.

We commence by laying the groundwork for item (1).

9.2.1. *Definition.* If $f \geq 0$ is a nonnegative measurable function (with respect to \mathcal{S}), we define

$$\int f d\mu = \begin{cases} +\infty & \text{if } f \notin \mathcal{L}^1(\mu) \\ \text{as usual} & \text{if } f \in \mathcal{L}^1(\mu). \end{cases}$$

The first properties of this notation are gathered in the following proposition:

9.2.2. *Proposition.* Let f, g and f_n ($n = 1, 2, 3, \dots$) be measurable functions ≥ 0 , and let c be a real number ≥ 0 . Then:

- (i) $\int c f d\mu = c \int f d\mu$.
- (ii) $\int (f + g) d\mu = \int f d\mu + \int g d\mu$.
- (iii) $f \leq g$ μ -a.e. $\Rightarrow \int f d\mu \leq \int g d\mu$.
- (iv) $f_n \uparrow f$ μ -a.e. $\Rightarrow \int f_n d\mu \uparrow \int f d\mu$.

Proof. (i) When $\int f d\mu = +\infty$, the convention $0 \cdot (+\infty) = 0$ saves the day (1.15.4).

(ii) Since $0 \leq f, g \leq f + g$,

$$\int (f+g) d\mu < +\infty \Leftrightarrow f+g \in \mathcal{L}^1 \Leftrightarrow f, g \in \mathcal{L}^1 \Leftrightarrow \int f d\mu + \int g d\mu < +\infty,$$

in which case the asserted equality is true by the additivity of integration (4.4.6). Otherwise, the equality reduces to $+\infty = +\infty$.

(iii) Assuming $f \leq g$ μ -a.e., we are to show that $\int f d\mu \leq \int g d\mu$. This is trivial if $\int g d\mu = +\infty$. Otherwise, $g \in \mathcal{L}^1(\mu)$, therefore $f \in \mathcal{L}^1(\mu)$ and $\int f d\mu \leq \int g d\mu$ by 4.4.19.

(iv) Assuming $f_n \uparrow f$ μ -a.e., we are to show that $\int f_n d\mu \uparrow \int f d\mu$. At any rate, $\int f_n d\mu \uparrow$ by (iii). If $\int f d\mu < +\infty$ then f and the f_n are μ -integrable and the assertion follows from the monotone convergence theorem (4.5.3). Otherwise $f \notin \mathcal{L}^1(\mu)$; it then follows from the monotone convergence theorem that either some f_n fails to be integrable, or every f_n is integrable but the sequence $\int f_n d\mu$ is unbounded, and in either case the assertion that $\sup \int f_n d\mu = \int f d\mu$ reduces to $+\infty = +\infty$. \diamond

The concept of indefinite integral extends to measurable functions ≥ 0 (and, with trivial modifications, to functions that are ≥ 0 μ -a.e., an extension for which we have no need):

9.2.3. *Definition.* If f is a measurable function ≥ 0 , a set function $f \cdot \mu : \mathcal{S} \rightarrow [0, +\infty]$ is defined by the formula

$$(f \cdot \mu)(E) = \int \varphi_E f d\mu \quad (E \in \mathcal{S}),$$

where the symbol on the right side, also written $\int_E f d\mu$, has the value assigned to it by Definition 9.2.1.

The properties of this notation are readily derived from Proposition 9.2.2:

9.2.4. *Proposition.* Let f, g and f_n ($n = 1, 2, 3, \dots$) be measurable functions ≥ 0 , and let c be a real number ≥ 0 . Then:

(1) $f \cdot \mu$ is a measure on \mathcal{S} such that

$$E \in \mathcal{S}, \mu(E) = 0 \Rightarrow (f \cdot \mu)(E) = 0.$$

$$(2) (cf) \cdot \mu = c(f \cdot \mu).$$

$$(3) (f + g) \cdot \mu = f \cdot \mu + g \cdot \mu.$$

$$(4) f \leq g \text{ } \mu\text{-a.e.} \Rightarrow f \cdot \mu \leq g \cdot \mu.$$

$$(5) f_n \uparrow f \text{ } \mu\text{-a.e.} \Rightarrow f_n \cdot \mu \uparrow f \cdot \mu \text{ on } \mathcal{S}.$$

$$(6) \varphi_F \cdot \mu = \mu_F \text{ for all } F \in \mathcal{S}.$$

$$(7) f \cdot \mu \text{ is a finite measure} \Leftrightarrow f \in \mathcal{L}^1(\mu).$$

$$(8) (fg) \cdot \mu = f \cdot (g \cdot \mu).$$

Proof. (1) We verify the criteria of Definition 2.4.12, by showing that the nonnegative function $f \cdot \mu$ vanishes at the empty set and is countably additive.

Since $\varphi_\emptyset f = 0$, we have $(f \cdot \mu)(\emptyset) = \int 0 d\mu = 0$.

If $E, F \in \mathcal{S}$ and $E \cap F = \emptyset$ then $\varphi_{E \cup F} f = \varphi_E f + \varphi_F f$, whence

$$(f \cdot \mu)(E \cup F) = (f \cdot \mu)(E) + (f \cdot \mu)(F)$$

by (ii) of 9.2.2, thus $f \cdot \mu$ is finitely additive. If (E_n) is a sequence of pairwise disjoint sets in \mathcal{S} with union E then, writing $F_n = \bigcup_{k=1}^n E_k$, we have $\varphi_{F_n} f \uparrow \varphi_E f$, consequently $(f \cdot \mu)(F_n) \uparrow (f \cdot \mu)(E)$ by (iv) of 9.2.2; since $f \cdot \mu$ is finitely additive, this means that

$$(f \cdot \mu)(E) = \lim_{n \rightarrow \infty} (f \cdot \mu)(F_n) = \lim_{n \rightarrow \infty} \sum_{k=1}^n (f \cdot \mu)(E_k),$$

thus f is countably additive.

Finally, if $\mu(E) = 0$ then $\varphi_E f = 0$ μ -a.e., thus $\varphi_E f$ is μ -integrable with integral 0, that is, $(f \cdot \mu)(E) = 0$.

(2) For all $E \in \mathcal{S}$, citing (i) of 9.2.2 at the appropriate step we have, for all $E \in \mathcal{S}$,

$$[(cf) \cdot \mu](E) = \int \varphi_E (cf) d\mu = \int c(\varphi_E f) d\mu = c \int \varphi_E f d\mu = c(f \cdot \mu)(E),$$

whence $(cf) \cdot \mu = c(f \cdot \mu)$.

(3) For all $E \in \mathcal{S}$,

$$\begin{aligned} [(f+g) \cdot \mu](E) &= \int \varphi_E(f+g) d\mu = \int (\varphi_E f + \varphi_E g) d\mu \\ &= \int \varphi_E f d\mu + \int \varphi_E g d\mu \\ &= (f \cdot \mu)(E) + (g \cdot \mu)(E) \end{aligned}$$

(the next-to-last equality by (ii) of 9.2.2), whence $(f+g) \cdot \mu = f \cdot \mu + g \cdot \mu$.

(4) For all $E \in \mathcal{S}$, $\varphi_E f \leq \varphi_E g$ μ -a.e., therefore $(f \cdot \mu)(E) \leq (g \cdot \mu)(E)$ by (iii) of 9.2.2.

(5) For all $E \in \mathcal{S}$, $\varphi_E f_n \uparrow \varphi_E f$ μ -a.e., therefore $(f_n \cdot \mu)(E) \uparrow (f \cdot \mu)(E)$ by (iv) of 9.2.2.

(6) If $F \in \mathcal{S}$ then, for all $E \in \mathcal{S}$,

$$\begin{aligned} (\varphi_F \cdot \mu)(E) &= \int \varphi_E \varphi_F d\mu = \int \varphi_{E \cap F} d\mu \\ &= \mu(E \cap F) = \mu_F(E) \end{aligned}$$

(the last equality by Definition 9.1.12), whence $\varphi_F \cdot \mu = \mu_F$.

(7) The assertion follows from the chain of equivalences

$$\begin{aligned} f \cdot \mu \text{ finite} &\Leftrightarrow (f \cdot \mu)(X) < +\infty \Leftrightarrow \int \varphi_X f d\mu < +\infty \\ &\Leftrightarrow \int f d\mu < +\infty \Leftrightarrow f \in \mathcal{L}^1(\mu). \end{aligned}$$

(8) If $F \in \mathcal{S}$ then, for all $E \in \mathcal{S}$,

$$\begin{aligned} [(\varphi_F g) \cdot \mu](E) &= \int \varphi_E (\varphi_F g) d\mu = \int \varphi_{E \cap F} g d\mu \\ &= (g \cdot \mu)(E \cap F) = (g \cdot \mu)_F(E), \end{aligned}$$

therefore $(\varphi_F g) \cdot \mu = (g \cdot \mu)_F = \varphi_F \cdot (g \cdot \mu)$ by (6). Thus, the equality (8) holds when f is the characteristic function of a measurable set; the case that f is a simple function then follows from (2) and (3). In general, let (f_n) be a sequence of simple functions such that $0 \leq f_n \uparrow f$. Then $f_n g \uparrow fg$, therefore $(f_n g) \cdot \mu \uparrow (fg) \cdot \mu$ by (5); but $(f_n g) \cdot \mu = f_n \cdot (g \cdot \mu)$ by the preceding case, and $f_n \cdot (g \cdot \mu) \uparrow f \cdot (g \cdot \mu)$, thus

$$f \cdot (g \cdot \mu) = \sup_n [f_n \cdot (g \cdot \mu)] = \sup_n [(f_n g) \cdot \mu] = (fg) \cdot \mu. \quad \diamond$$

The foregoing item (8) is the key to the relation between the integration theories for μ and $g \cdot \mu$:

9.2.5. Theorem. Let $g \geq 0$ be measurable and let $f : X \rightarrow \mathbb{C}$. The following conditions are equivalent:

- (a) $f \in \mathcal{L}^1_{\mathbb{C}}(g \cdot \mu)$;
- (b) f is measurable and $fg \in \mathcal{L}^1_{\mathbb{C}}(\mu)$.

In this case,

$$\int f d(g \cdot \mu) = \int fg d\mu.$$

Proof. Both (a) and (b) entail the measurability of f (with respect to \mathcal{S}), so let us assume f measurable at the outset. Since g is real-valued, we have

$$\operatorname{Re}(fg) = (\operatorname{Re} f)g, \quad \operatorname{Im}(fg) = (\operatorname{Im} f)g,$$

so we can suppose that f is real-valued (6.4.4). In fact, we can suppose further that $f \geq 0$; for, if the equivalence (a) \Leftrightarrow (b) is valid when $f \geq 0$, then its validity for real-valued f follows from the formulas

$$f = f^+ - f^-, \quad (fg)^+ = f^+g, \quad (fg)^- = f^-g$$

and the fact that f is $g \cdot \mu$ -integrable if and only if f^+ and f^- are, whereas fg is μ -integrable if and only if $(fg)^+$ and $(fg)^-$ are (4.4.12).

Assuming $f \geq 0$, we have

$$\begin{aligned} f \in \mathcal{L}^1(g \cdot \mu) &\Leftrightarrow f \cdot (g \cdot \mu) \text{ is finite} \\ &\Leftrightarrow (fg) \cdot \mu \text{ is finite} \\ &\Leftrightarrow fg \in \mathcal{L}^1(\mu) \end{aligned}$$

by items (7) and (8) of 9.2.4, and in this case

$$\begin{aligned} \int f d(g \cdot \mu) &= \int \varphi_X f d(g \cdot \mu) = [f \cdot (g \cdot \mu)](X) \\ [(fg) \cdot \mu](X) &= \int \varphi_X (fg) d\mu = \int fg d\mu. \quad \diamond \end{aligned}$$

Next, we relax the requirement that the function f in the expression $f \cdot \mu$ is positive; as in the preceding section (cf. 9.1.3), we are willing to deal with $+\infty$ and $-\infty$, but we don't want them on our plate at the same time:

9.2.6. *Definition.* We write \mathcal{D} for the set of all functions

$$f - g,$$

where f and g are measurable functions ≥ 0 and at least one of f, g is μ -integrable (so that the difference $\int f d\mu - \int g d\mu$ is defined in $\overline{\mathbb{R}}$). When it is necessary to highlight the dependence of \mathcal{D} on the measure μ , we write $\mathcal{D}(\mu)$.

9.2.7. *Remarks.* The set \mathcal{D} just defined has the following properties:

- (i) $\mathcal{L}^1(\mu) \subset \mathcal{D}$ (because $f = f^+ - f^-$; cf. 4.4.12).
- (ii) $c \in \mathbb{R}, h \in \mathcal{D} \Rightarrow ch \in \mathcal{D}$.

(iii) If f is measurable and either $f \geq 0$ or $f \leq 0$, then $f \in \mathcal{D}$ (because $0 \in \mathcal{L}^1$ and $f = f - 0 = 0 - (-f)$).

In general \mathcal{D} is not closed under addition (Exercise 1).

Suppose $h \in \mathcal{D}$ and write $h = f - g$ as in Definition 9.2.6. Since one of f, g is μ -integrable, one of the measures $f \cdot \mu, g \cdot \mu$ is finite by (7) of 9.2.4, therefore the difference $f \cdot \mu - g \cdot \mu$ is defined and is a signed measure (9.1.3, 9.1.4). This signed measure depends only on h and not on the particular representation of h in the form $f - g$. For, suppose also $h = f_1 - g_1$, where f_1, g_1 are measurable functions ≥ 0 and at least one of them is μ -integrable. Then

$$(*) \quad f + g_1 = f_1 + g,$$

therefore

$$(**) \quad f \cdot \mu + g_1 \cdot \mu = f_1 \cdot \mu + g \cdot \mu$$

by (3) of 9.2.4. We note that if f is integrable then so is f_1 ; for, if f_1 were not integrable then g_1 would be integrable, whence the absurdity that the left side of (*) is integrable but the right side is not. Similarly, if g is integrable then so is g_1 . If g (hence also g_1) is integrable then transposition of the finite measures $g \cdot \mu$ and $g_1 \cdot \mu$ in (**) yields the desired equality

$$f \cdot \mu - g \cdot \mu = f_1 \cdot \mu - g_1 \cdot \mu.$$

On the other hand if g is not integrable then f is integrable, the foregoing reasoning applied to $-h$ yields

$$g \cdot \mu - f \cdot \mu = g_1 \cdot \mu - f_1 \cdot \mu,$$

and the desired equality results on multiplication by -1 .

9.2.8. *Definition.* If $h \in \mathcal{D}$ and $h = f - g$ as in 9.2.6 then $f \cdot \mu - g \cdot \mu$ is a signed measure (9.1.3) and, by the preceding remarks, it depends only on h and not on the particular representation of h as such a difference; we define

$$h \cdot \mu = f \cdot \mu - g \cdot \mu.$$

Thus, for all $E \in \mathcal{S}$,

$$\begin{aligned} (h \cdot \mu)(E) &= (f \cdot \mu)(E) - (g \cdot \mu)(E) \\ &= \int \varphi_E f d\mu - \int \varphi_E g d\mu \\ &= \int_E f d\mu - \int_E g d\mu. \end{aligned}$$

In particular,

$$(h \cdot \mu)(X) = \int f d\mu - \int g d\mu,$$

so there is no ambiguity in defining

$$\int h d\mu = \int f d\mu - \int g d\mu.$$

9.2.9. *Remark.* With notations as in the preceding definition,

$$(h \cdot \mu)(E) = \int \varphi_E h d\mu \quad \text{for all } E \in \mathcal{S}.$$

{Proof: Write $h = f - g$ as in 9.2.6 and let $E \in \mathcal{S}$. We have $\varphi_E h = \varphi_E f - \varphi_E g$, where $\varphi_E f, \varphi_E g$ are measurable functions ≥ 0 , at least one of which is μ -integrable, consequently $\varphi_E h \in \mathcal{D}$ and

$$\int \varphi_E h d\mu = \int \varphi_E f d\mu - \int \varphi_E g d\mu$$

by the definition of the left side (9.2.8), that is, $\int \varphi_E h d\mu = (h \cdot \mu)(E)$. This expression may also be written $\int_E h d\mu$, extending the notation in 9.2.3.}

The functions in \mathcal{D} are conveniently characterized as follows:

9.2.10. Proposition. *The following conditions on a function $h : X \rightarrow \mathbb{R}$ are equivalent:*

- (a) $h \in \mathcal{D}$;
- (b) h is measurable and at least one of h^+, h^- is μ -integrable.

Proof. (b) \Rightarrow (a): Immediate from $h = h^+ - h^-$.

(a) \Rightarrow (b): Write $h = f - g$ as in 9.2.6. At any rate, h is measurable and we know that either f or g is μ -integrable.

case 1: $f \in \mathcal{L}^1(\mu)$.

Let $A = \{x : h(x) \geq 0\} = \{x : f(x) \geq g(x)\}$. Then $\varphi_A h = h^+$ and

$$0 \leq h^+ = \varphi_A h = \varphi_A(f - g) = \varphi_A f - \varphi_A g \leq \varphi_A f \in \mathcal{L}^1(\mu),$$

therefore $h^+ \in \mathcal{L}^1(\mu)$.

case 2: $g \in \mathcal{L}^1(\mu)$.

Then $-h = g - f$, so by the proof of case 1, $(-h)^+ \in \mathcal{L}^1(\mu)$, that is, $h^- \in \mathcal{L}^1(\mu)$. \diamond

For the signed measures $h \cdot \mu$ ($h \in \mathcal{D}$) of the present section, the Hahn-Jordan decomposition and total variation defined in the preceding section are readily expressed in terms of h :

9.2.11. Corollary. *For every $h \in \mathcal{D}$,*

$$(h \cdot \mu)^+ = h^+ \cdot \mu, \quad (h \cdot \mu)^- = h^- \cdot \mu, \quad |h \cdot \mu| = |h| \cdot \mu.$$

Proof. We recall that $|h \cdot \mu| = (h \cdot \mu)^+ + (h \cdot \mu)^-$ (9.1.22); on the other hand, $|h|$ is a measurable function ≥ 0 , thus $|h| \cdot \mu$ is the measure defined in 9.2.3. If

$$A = \{x : h(x) \geq 0\},$$

then $\varphi_A h = h^+$. Let

$$\begin{aligned} B &= X - A = \{x : h(x) < 0\}, \\ C &= \{x : h(x) \leq 0\} = \{x : (-h)(x) \geq 0\}; \end{aligned}$$

then $\varphi_C(-h) = (-h)^+ = h^-$, thus

$$h^- = -\varphi_C h = -\varphi_B h$$

(note that the latter equality holds trivially at the points where h is 0). Summarizing,

$$\varphi_A h = h^+, \quad \varphi_B h = -h^-.$$

Now, $h = h^+ - h^-$, where one of h^+, h^- is μ -integrable by the preceding proposition, therefore

$$h \cdot \mu = h^+ \cdot \mu - h^- \cdot \mu$$

by Definition 9.2.8. If $(h \cdot \mu)_A$ is the signed measure defined in 9.1.12 then, for all $E \in \mathcal{S}$,

$$\begin{aligned} (h \cdot \mu)_A(E) &= (h \cdot \mu)(A \cap E) && \text{(by 9.1.12)} \\ &= \int \varphi_{A \cap E} h d\mu && \text{(by 9.2.9)} \\ &= \int \varphi_E \varphi_A h d\mu \\ &= \int \varphi_E h^+ d\mu \\ &= (h^+ \cdot \mu)(E) && \text{(by 9.2.3)} \end{aligned}$$

thus $(h \cdot \mu)_A = h^+ \cdot \mu \geq 0$. Similarly, for all $E \in \mathcal{S}$,

$$\begin{aligned} (h \cdot \mu)_B &= (h \cdot \mu)(B \cap E) = \int \varphi_{B \cap E} h d\mu \\ &= \int \varphi_E \varphi_B h d\mu = \int \varphi_E (-h^-) d\mu \\ &= \int (0 - \varphi_E h^-) d\mu \\ &= \int 0 d\mu - \int \varphi_E h^- d\mu && \text{(by 9.2.8)} \\ &= -(h^- \cdot \mu)(E), \end{aligned}$$

thus $(h \cdot \mu)_B = -(h^- \cdot \mu) \leq 0$. In other words, $A \geq 0$ and $X - A \leq 0$ with respect to the signed measure $h \cdot \mu$ (9.1.16); thus A and $X - A$ define a Jordan-Hahn decomposition (9.1.22) of $h \cdot \mu$, with

$$\begin{aligned} (h \cdot \mu)^+ &= (h \cdot \mu)_A = h^+ \cdot \mu, \\ (h \cdot \mu)^- &= -(h \cdot \mu)_{X-A} = h^- \cdot \mu, \end{aligned}$$

consequently, citing (3) of 9.2.4 at the appropriate step,

$$\begin{aligned} |h \cdot \mu| &= (h \cdot \mu)^+ + (h \cdot \mu)^- = h^+ \cdot \mu + h^- \cdot \mu \\ &= (h^+ + h^-) \cdot \mu = |h| \cdot \mu. \quad \diamond \end{aligned}$$

9.2.12. *Definition.* A signed measure ν on \mathcal{S} is said to be **absolutely continuous** with respect to the measure μ , written

$$\nu \ll \mu,$$

if it satisfies the condition

$$E \in \mathcal{S}, \mu(E) = 0 \Rightarrow \nu(E) = 0.$$

(The case that ν is a finite signed measure—that is, a real measure—was considered in 4.8.6.) For example, if f is a measurable function ≥ 0 , then $f \cdot \mu \ll \mu$ by (1) of 9.2.4.

9.2.13. *Proposition.* For a signed measure ν on \mathcal{S} , the following conditions are equivalent:

- (a) $\nu \ll \mu$;
- (a') $E \in \mathcal{S}, \mu(E) = 0 \Rightarrow \nu_E = 0$;
- (b) $|\nu| \ll \mu$;
- (c) $\nu^+ \ll \mu$ and $\nu^- \ll \mu$.

Proof. (a) \Rightarrow (a'): If $\mu(E) = 0$ then, for all $F \in \mathcal{S}$, $\mu(E \cap F) = 0$, therefore $\nu(E \cap F) = 0$ by (a); thus $\nu_E = 0$.

(a') \Rightarrow (c): Let $A, X - A$ be a Hahn decomposition for ν (9.1.22); thus $A \in \mathcal{S}$ and

$$\nu^+ = \nu_A, \quad \nu^- = -\nu_{X-A}.$$

If $\mu(E) = 0$ then $\nu_E = 0$ by (a'), therefore by 9.1.23 we have

$$(\nu^+)_E = (\nu_E)^+ = 0^+ = 0$$

and similarly $(\nu^-)_E = 0$; thus $\nu^+(E) = 0$ and $\nu^-(E) = 0$, and we have shown that $\nu^+ \ll \mu$ and $\nu^- \ll \mu$.

(c) \Rightarrow (a): This is clear from the formula $\nu(E) = \nu^+(E) - \nu^-(E)$ ($E \in \mathcal{S}$).

Summarizing: (a) \Leftrightarrow (a') \Leftrightarrow (c). Finally,

(b) \Leftrightarrow (c): Since $|\nu| = \nu^+ + \nu^-$ (9.1.22) we have

$$|\nu|(E) = \nu^+(E) + \nu^-(E) \quad \text{for all } E \in \mathcal{S}.$$

By the positivity of $|\nu|$, ν^+ and ν^- , it follows that

$$|\nu|(E) = 0 \Leftrightarrow \nu^+(E) = 0 \text{ \& } \nu^-(E) = 0,$$

whence the equivalence of (b) and (c). \diamond

The signed measures $h \cdot \mu$ ($h \in \mathcal{D}$) are all absolutely continuous with respect to μ :

9.2.14. Theorem. Let $h \in \mathcal{D}$. Then:

- (i) $h \cdot \mu \ll \mu$.
- (ii) If μ is σ -finite, then so is $|h \cdot \mu|$.

Proof. (i) Since $|h \cdot \mu| = |h| \cdot \mu$ (9.2.11) and $|h| \cdot \mu \ll \mu$ by (1) of 9.2.4, we have $h \cdot \mu \ll \mu$ by the preceding proposition.

(ii) Since $|h \cdot \mu| = |h| \cdot \mu$ we can suppose that $h \geq 0$, in which case $h \cdot \mu$ is a measure. Assuming μ σ -finite, there exists a sequence (P_n) of sets in \mathcal{S} such that $P_n \uparrow X$ and $\mu(P_n) < +\infty$ for all n . Let

$$Q_n = \{x : h(x) \leq n\};$$

then $Q_n \uparrow X$, therefore $P_n \cap Q_n \uparrow X$, where $\mu(P_n \cap Q_n) < +\infty$ for all n . It will suffice to show that $(h \cdot \mu)(P_n \cap Q_n) < +\infty$ for all n ; indeed,

$$0 \leq h\varphi_{P_n \cap Q_n} \leq n\varphi_{P_n \cap Q_n} \in \mathcal{L}^1(\mu),$$

therefore $h\varphi_{P_n \cap Q_n} \in \mathcal{L}^1(\mu)$, thus

$$(h \cdot \mu)(P_n \cap Q_n) = \int \varphi_{P_n \cap Q_n} h d\mu < +\infty. \diamond$$

The preceding theorem says that if μ is a σ -finite measure then the signed measures $h \cdot \mu$ ($h \in \mathcal{D}$) are absolutely continuous with respect to μ and their total variations are σ -finite. The big result goes in the reverse direction:

9.2.15. Theorem. (**Radon–Nikodým**)¹ If μ is a σ -finite measure on \mathcal{S} and if ν is a signed measure on \mathcal{S} such that $\nu \ll \mu$ and $|\nu|$ is σ -finite, then $\nu = h \cdot \mu$ for some $h \in \mathcal{D}$.

Proof. Suppose first that $\nu \geq 0$, that is, ν is a measure. Since both μ and ν are σ -finite, there exists a sequence (P_n) of pairwise disjoint sets in \mathcal{S} such that

$$X = \bigcup_{n=1}^{\infty} P_n, \quad \mu(P_n) < +\infty, \quad \nu(P_n) < +\infty.$$

(Let (Q_n) be a sequence of pairwise disjoint measurable sets with union X and $\mu(Q_n) < +\infty$ for all n , let (R_n) be a similar sequence for ν , and let (P_n) be an enumeration of the pairwise disjoint sets $Q_m \cap R_n$.) The measures $\varphi_{P_n} \cdot \nu = \nu_{P_n}$ and $\varphi_{P_n} \cdot \mu = \mu_{P_n}$ are finite. Moreover, for each n ,

$$\nu_{P_n} \ll \mu_{P_n};$$

for, if $E \in \mathcal{S}$ with $\mu_{P_n}(E) = 0$, that is, $\mu(P_n \cap E) = 0$, then $\nu(P_n \cap E) = 0$ by the assumption $\nu \ll \mu$, whence $\nu_{P_n}(E) = 0$. By the Radon–Nikodým theorem for finite measures (4.8.11) there exists, for each

¹ Johann Karl August Radon (1887–1956), Otton Martin Nikodým (1887–1974).

index n , a measurable function $h_n \geq 0$ such that

$$\nu_{P_n} = h_n \cdot \mu_{P_n},$$

that is,

$$\nu_{P_n} = h_n \cdot (\varphi_{P_n} \cdot \mu) = (h_n \varphi_{P_n}) \cdot \mu$$

by (8) of 9.2.4. Replacing h_n by $h_n \varphi_{P_n}$, we can suppose that

$$\varphi_{P_n} h_n = h_n,$$

that is, $h_n = 0$ on $X - P_n$, and

$$\varphi_{P_n} \cdot \nu = h_n \cdot \mu \quad \text{for all } n.$$

Define a function $h : X \rightarrow [0, +\infty)$ as the pointwise sum

$$h = \sum_{n=1}^{\infty} h_n$$

(note that for each x , $h_n(x)$ is nonzero for at most one value of n). Since

$$\sum_{k=1}^n h_k \uparrow h \quad \text{pointwise on } X,$$

h is measurable with respect to \mathcal{S} ; also $h \geq 0$, so $h \in \mathcal{D}$ by item (iii) of 9.2.7. Writing

$$U_n = \bigcup_{k=1}^n P_k,$$

we have $U_n \uparrow X$ and

$$\begin{aligned} (*) \quad \varphi_{U_n} \cdot \nu &= \left(\sum_{k=1}^n \varphi_{P_k} \right) \cdot \nu = \sum_{k=1}^n \varphi_{P_k} \cdot \nu \\ &= \sum_{k=1}^n h_k \cdot \mu = \left(\sum_{k=1}^n h_k \right) \cdot \mu; \end{aligned}$$

since $\varphi_{U_n} \uparrow \varphi_X = 1$ and $\sum_{k=1}^n h_k \uparrow h$, it follows from (*) and item (5) of 9.2.4 that $1 \cdot \nu = h \cdot \mu$, that is, $\nu = h \cdot \mu$.

Now consider the general case (ν a signed measure not necessarily ≥ 0). By the Jordan-Hahn decomposition theorem of the preceding section (9.1.20, 9.1.22), we have $\nu = \nu^+ - \nu^-$, where at least one of the measures ν^+ , ν^- is finite, and $\nu^+ \ll \mu$ and $\nu^- \ll \mu$ by 9.2.13. Moreover, it is clear from the inequalities $0 \leq \nu^+$, $\nu^- \leq |\nu|$ that ν^+ and ν^- are also σ -finite. By the case proved in the preceding paragraph, there exist measurable functions $f, g \geq 0$ such that $\nu^+ = f \cdot \mu$ and $\nu^- = g \cdot \mu$; moreover, at least one of f, g is μ -integrable by (7) of 9.2.4, therefore the function $h = f - g$ belongs to \mathcal{D} (9.2.6) and, by Definition 9.2.8,

$$h \cdot \mu = f \cdot \mu - g \cdot \mu = \nu^+ - \nu^- = \nu. \quad \diamond$$

An alternative proof of the Radon-Nikodym theorem (see Exercise 5) can be based on the following extension of (8) of 9.2.4:

9.2.16. Proposition. Let g be a measurable function ≥ 0 (so that $g \cdot \mu$ is a measure) and let $f \in \mathcal{D}(g \cdot \mu)$ (so that the signed measure $f \cdot (g \cdot \mu)$ is defined). Then $fg \in \mathcal{D}(\mu)$ and

$$(fg) \cdot \mu = f \cdot (g \cdot \mu).$$

Proof. Since $g \cdot \mu$ is a measure (9.2.4) and $f \in \mathcal{D}(g \cdot \mu)$, we know from Definitions 9.2.6 and 9.2.8 that $f = u - v$ with u and v measurable functions ≥ 0 , at least one of them $(g \cdot \mu)$ -integrable, and that

$$\begin{aligned} f \cdot (g \cdot \mu) &= u \cdot (g \cdot \mu) - v \cdot (g \cdot \mu) \\ &= (ug) \cdot \mu - (vg) \cdot \mu \end{aligned}$$

(the latter equality by (8) of 9.2.4). Moreover, since one of u, v is $(g \cdot \mu)$ -integrable, it follows from Theorem 9.2.5 that one of the (nonnegative, measurable) functions ug, vg is μ -integrable, therefore $ug - vg \in \mathcal{D}(\mu)$ and

$$(ug - vg) \cdot \mu = (ug) \cdot \mu - (vg) \cdot \mu,$$

that is, $(fg) \cdot \mu = f \cdot (g \cdot \mu)$. \diamond

Exercises

1. If there exists a measurable function $h : X \rightarrow \mathbb{R}$ such that neither h^+ nor h^- is μ -integrable, then h^+ and $-h^-$ belong to \mathcal{D} but their sum does not (9.2.10). For example, if there exists a set $A \in \mathcal{S}$ such that $\mu(A) = \mu(X - A) = +\infty$, then $h = \varphi_A - \varphi_{X-A}$ is such a function.

2. The conclusion of the Radon-Nikodym theorem (9.2.15) may fail if the assumption that $|\nu|$ is σ -finite is omitted.

{Hint: Let X be a singleton and let μ, ν be the measures on $\mathcal{P}(X) = \{\emptyset, X\}$ such that $\mu(X) = 1$ and $\nu(X) = +\infty$. It is trivial that $\nu \ll \mu$ whereas $\mathcal{D}(\mu)$ is the set of constant functions on X .}

3. (i) If μ is σ -finite and h is a measurable function on X , then there exists a sequence (E_n) of measurable sets such that $E_n \uparrow X$, $\mu(E_n) < +\infty$ and $\varphi_{E_n} h$ is μ -integrable for all n .

{Hint: Let (P_n) be a sequence in \mathcal{S} such that $P_n \uparrow X$ and $\mu(P_n) < +\infty$ for all n , let

$$Q_n = \{x : |h(x)| \leq n\}$$

and contemplate $E_n = P_n \cap Q_n$ }

(ii) If $h_1, h_2 \in \mathcal{D}$ and $h_1 \leq h_2$ μ -a.e., then $h_1 \cdot \mu \leq h_2 \cdot \mu$.

{Hint: Since $\varphi_E h_1 \leq \varphi_E h_2$ μ -a.e. for all $E \in \mathcal{S}$, it suffices, in view of 9.2.9, to show that $\int h_1 d\mu \leq \int h_2 d\mu$. Write $h_1 = f_1 - g_1$ and $h_2 = f_2 - g_2$

as in Definition 9.2.6. Argue that if $g_1 \in \mathcal{L}^1(\mu)$ then also $g_2 \in \mathcal{L}^1(\mu)$, whereas if $g_1 \notin \mathcal{L}^1(\mu)$ then the left member of the asserted inequality is $-\infty$.}

(iii) Suppose μ is σ -finite and let $h_1, h_2 \in \mathcal{D}$. Then

$$h_1 \cdot \mu \leq h_2 \cdot \mu \quad \Leftrightarrow \quad h_1 \leq h_2 \quad \mu\text{-a.e.}$$

$$h_1 \cdot \mu = h_2 \cdot \mu \quad \Leftrightarrow \quad h_1 = h_2 \quad \mu\text{-a.e.}$$

{Hint: It clearly suffices to prove the first equivalence. In view of (ii), it suffices to show that $h_1 \cdot \mu \leq h_2 \cdot \mu \Rightarrow h_1 \leq h_2$ μ -a.e. By (i) there exists a sequence (E_n) of measurable sets such that $E_n \uparrow X$ and $\varphi_{E_n} h_1, \varphi_{E_n} h_2$ are μ -integrable for all n . Note that $(\varphi_{E_n} h_1) \cdot \mu \leq (\varphi_{E_n} h_2) \cdot \mu$ and cite (iv) of 4.7.2.}

4. (*Polar decomposition*) If ν is any signed measure on \mathcal{S} , then there exists a function $f \in \mathcal{D}(|\nu|)$ such that $\nu = f \cdot |\nu|$ and $|f| = 1$.

{Hint: Choose $A \in \mathcal{S}$ such that $\nu^+ = \nu_A$ and $\nu^- = -\nu_{X-A}$ (9.1.22) and contemplate $f = \varphi_A - \varphi_{X-A}$.}

5. In the proof of the Radon–Nikodym theorem (9.2.15) the general case can be derived from the special case of a measure by means of Exercise 4.

{Hint: Write $\nu = f \cdot |\nu|$ as in Exercise 4 and apply the first part of the proof of 9.2.15 to write $|\nu| = g \cdot \mu$ with $g \geq 0$. Look at $\nu = f \cdot |\nu| = f \cdot (g \cdot \mu)$ in the light of Proposition 9.2.16.}

9.3. Lebesgue Decomposition of Measures

Throughout this section, (X, \mathcal{S}) is a fixed measurable space; all measures and signed measures under consideration are defined on \mathcal{S} .

The topic taken up in this section is an analogue, for measures, of the decomposition of a function of bounded variation as the sum of an absolutely continuous function and a singular function (§5.12).

9.3.1. *Definition.* A pair μ, ν of signed measures on \mathcal{S} are said to be **mutually singular**, written

$$\mu \perp \nu,$$

if there exists a set $A \in \mathcal{S}$ such that

$$\mu_A = \mu \quad \text{and} \quad \nu_{X-A} = \nu.$$

The first properties of this notation are as follows:

9.3.2. Proposition. Let μ and ν be signed measures on \mathcal{S} , and let $E \in \mathcal{S}$.

(i) $\mu \perp \nu \Leftrightarrow \nu \perp \mu$.

- (ii) $\mu_E = \mu \Leftrightarrow \mu_{X-E} = 0$.
- (iii) $\mu \perp \nu \Leftrightarrow \exists A \in \mathcal{S} \ni \mu_A = \mu \text{ \& } \nu_A = 0$.
- (iv) $\mu \perp \nu \Leftrightarrow \exists A, B \in \mathcal{S} \ni \mu_A = \mu, \nu_B = \nu, A \cap B = \emptyset$.
- (v) $\mu_E = 0 \Leftrightarrow |\mu|_E = 0$.
- (vi) $\mu_E = \mu \Leftrightarrow |\mu|_E = |\mu|$.
- (vii) $\mu \perp \nu \Leftrightarrow |\mu| \perp |\nu|$.
- (viii) $\mu \perp \mu \Leftrightarrow \mu = 0$.

Proof. For brevity we write $\mathbb{C}E = X - E$.

(i) Clear from $\mathbb{C}(\mathbb{C}E) = E$.

(ii) If $\mu_E = \mu$ then

$$\mu_{\mathbb{C}E} = (\mu_E)_{\mathbb{C}E} = \mu_{E \cap \mathbb{C}E} = \mu_{\emptyset} = 0.$$

Conversely, if $\mu_{\mathbb{C}E} = 0$ then, citing (v) of 9.1.13, we have

$$\mu = \mu_X = \mu_{E \cup \mathbb{C}E} = \mu_E + \mu_{\mathbb{C}E} = \mu_E + 0 = \mu_E.$$

(iii) By (ii), $\nu_A = 0 \Leftrightarrow \nu_{X-A} = \nu$, thus (iii) is equivalent to the condition defining $\mu \perp \nu$ (9.3.1).

(iv) If $\mu \perp \nu$ and if $A \in \mathcal{S}$ is chosen as in Definition 9.3.1, then the pair $A, B = X - A$ meets the requirements of the condition on the right. Conversely, if A and B meet the requirements of that condition, so that in particular $B \subset X - A$, then

$$\nu_{X-A} = (\nu_B)_{X-A} = \nu_{B \cap (X-A)} = \nu_B = \nu,$$

thus A satisfies the condition defining $\mu \perp \nu$ (9.3.1).

(v) By Proposition 9.1.23, $|\mu|_E = |\mu_E|$ and $|\mu_E| = 0 \Leftrightarrow \mu_E = 0$.

(vi) If $\mu_E = \mu$ then, by 9.1.23, $|\mu|_E = |\mu_E| = |\mu|$. Conversely, if $|\mu|_E = |\mu|$ then $|\mu|_{X-E} = 0$ by (ii), that is, $|\mu_{X-E}| = 0$, whence $\mu_{X-E} = 0$, and finally $\mu_E = \mu$ by (ii).

(vii) If $A \in \mathcal{S}$ then, by (vi),

$$\mu_A = \mu \text{ \& } \nu_{X-A} = \nu \Leftrightarrow |\mu|_A = |\mu| \text{ \& } |\nu|_{X-A} = |\nu|,$$

thus $\mu \perp \nu \Leftrightarrow |\mu| \perp |\nu|$ by Definition 9.3.1.

(viii) If $\mu \perp \mu$ and if $A \in \mathcal{S}$ is chosen as in 9.3.1, so that $\mu_A = \mu$ and $\mu_{X-A} = \mu$, then, by (ii), $0 = \mu_{X-A} = \mu$. Conversely, if $\mu = 0$ then every set in \mathcal{S} can play the role of A in 9.3.1 in showing that $\mu \perp \mu$. \diamond

If μ is a measure and ν is a signed measure, the relation $\nu \ll \mu$ is expressible in terms of null sets (9.2.12): $\mu(E) = 0 \Rightarrow \nu(E) = 0$. When μ and ν are both signed measures, the useful concept is as follows:

9.3.3. *Definition.* Let μ and ν be signed measures on \mathcal{S} . We say that ν is **absolutely continuous** with respect to μ , written

$$\nu \ll \mu,$$

if $\nu \ll |\mu|$ in the sense of 9.2.12, that is, if

$$E \in \mathcal{S}, |\mu|(E) = 0 \Rightarrow \nu(E) = 0.$$

9.3.4. *Remarks.* With notations as in 9.3.3, it follows from 9.2.13 that

$$\nu \ll \mu \Leftrightarrow |\nu| \ll |\mu| \Leftrightarrow \nu^+ \ll |\mu| \text{ \& } \nu^- \ll |\mu|,$$

therefore

$$\nu \ll \mu \Leftrightarrow \nu^+ \ll \mu \text{ \& } \nu^- \ll \mu.$$

Also, since $|\mu|$ is a measure, $|\mu|(E) = 0$ is equivalent to $|\mu|_E = 0$, in other words (9.1.23), to $|\mu_E| = 0$ and hence to $\mu_E = 0$. Thus the relation $\nu \ll \mu$ —equivalently $|\nu| \ll |\mu|$ —can also be expressed by the implication

$$E \in \mathcal{S}, \mu_E = 0 \Rightarrow \nu_E = 0.$$

9.3.5 CAUTION. if ν and μ are signed measures on \mathcal{S} such that $\nu \ll \mu$, it does not follow that $\mu(E) = 0 \Rightarrow \nu(E) = 0$ (Exercise 1).

9.3.6. *Definition.* Let μ and ν be signed measures on \mathcal{S} . A **Lebesgue decomposition** of ν with respect to μ is a pair ν_1, ν_2 of signed measures on \mathcal{S} such that

$$\nu = \nu_1 + \nu_2, \quad \nu_1 \ll \mu \quad \text{and} \quad \nu_2 \perp \mu.$$

(In particular, the sum $\nu_1 + \nu_2$ is defined, and $|\nu_1| \ll |\mu|$, $|\nu_2| \perp |\mu|$ by 9.3.4 and (vii) of 9.3.2.) One calls ν_1 the *absolutely continuous part* of ν , and ν_2 the *singular part* of ν , with respect to μ .

Our goal in this section is to show that if μ and ν are signed measures on \mathcal{S} such that $|\mu|$ and $|\nu|$ are σ -finite, then there exists a Lebesgue decomposition of ν with respect to μ (and, of course, of μ with respect to ν , by the symmetry of the hypotheses). First we look at the question of uniqueness of such a decomposition:

9.3.7. *Proposition.* Let μ and ν be signed measures on \mathcal{S} and suppose

$$\nu = \nu_1 + \nu_2 = \rho_1 + \rho_2$$

are two Lebesgue decompositions of ν with respect to μ (9.3.6). Then:

(1) $\nu_2 = \rho_2$.

(2) If, moreover, $|\nu_2|$ is σ -finite, then $\nu_1 = \rho_1$.

Proof. By assumption $\nu_1 \ll \mu$, $\nu_2 \perp \mu$ and $\rho_1 \ll \mu$, $\rho_2 \perp \mu$. From Definition 9.3.3 we know that $\nu_1 \ll |\mu|$ and $\rho_1 \ll |\mu|$; also, by (vii) of 9.3.2, $\nu_2 \perp |\mu|$ and $\rho_2 \perp |\mu|$. Replacing μ by $|\mu|$, we can therefore suppose that μ is a measure.

(1) Since $\nu_2 \perp \mu$, by (iii) of 9.3.2 there exists a set $A \in \mathcal{S}$ such that

$$(\nu_2)_A = \nu_2, \quad \mu_A = 0 \quad (\text{that is, } \mu(A) = 0),$$

and similarly there exists a set $B \in \mathcal{S}$ such that

$$(\rho_2)_B = \rho_2, \quad \mu_B = 0 \quad (\text{that is, } \mu(B) = 0).$$

Then $\mu(A \cup B) \leq \mu(A) + \mu(B) = 0$, thus $\mu(A \cup B) = 0$; since $\nu_1 \ll \mu$, it follows from (a') of 9.2.13 that $(\nu_1)_{A \cup B} = 0$, therefore (9.1.13)

$$\begin{aligned} \nu_{A \cup B} &= (\nu_1 + \nu_2)_{A \cup B} = (\nu_1)_{A \cup B} + (\nu_2)_{A \cup B} \\ &= 0 + (\nu_2)_{A \cup B} = ((\nu_2)_A)_{A \cup B} \\ &= (\nu_2)_{A \cap (A \cup B)} = (\nu_2)_A = \nu_2. \end{aligned}$$

Similarly $\nu_{A \cup B} = \rho_2$, thus $\nu_2 = \nu_{A \cup B} = \rho_2$.

(2) By (1), we can write

$$\nu = \nu_1 + \nu_2 = \rho_1 + \nu_2;$$

assuming $|\nu_2|$ is σ -finite, our task is to 'cancel' ν_2 in the second equation. Let (E_n) be a sequence of pairwise disjoint sets in \mathcal{S} such that

$$X = \bigcup_{n=1}^{\infty} E_n, \quad |\nu_2|(E_n) < +\infty \quad \text{for all } n.$$

Then, for each n , the measure $|(\nu_2)_{E_n}| = |\nu_2|_{E_n}$ is finite, therefore so are the measures $(\nu_2)_{E_n}^+$, $(\nu_2)_{E_n}^-$, consequently $(\nu_2)_{E_n}$ is finite-valued (i.e., is a real measure). Now,

$$\nu_{E_n} = (\nu_1 + \nu_2)_{E_n} = (\nu_1)_{E_n} + (\nu_2)_{E_n},$$

and similarly

$$\nu_{E_n} = (\rho_1 + \nu_2)_{E_n} = (\rho_1)_{E_n} + (\nu_2)_{E_n},$$

thus

$$(\nu_1)_{E_n} + (\nu_2)_{E_n} = (\rho_1)_{E_n} + (\nu_2)_{E_n}$$

for all n ; since $(\nu_2)_{E_n}$ is finite-valued, it can be canceled in the preceding equation, so that

$$(\nu_1)_{E_n} = (\rho_1)_{E_n} \quad \text{for all } n.$$

Then, for each $F \in \mathcal{S}$, the sets $E_n \cap F$ are pairwise disjoint with union F , consequently

$$\begin{aligned} \nu_1(F) &= \sum_{n=1}^{\infty} \nu_1(E_n \cap F) = \sum_{n=1}^{\infty} (\nu_1)_{E_n}(F) \\ &= \sum_{n=1}^{\infty} (\rho_1)_{E_n}(F) = \sum_{n=1}^{\infty} \rho_1(E_n \cap F) = \rho_1(F), \end{aligned}$$

thus $\nu_1 = \rho_1$. \diamond

9.3.8. *Definition.* Signed measures μ, ν on \mathcal{S} are said to be **equivalent**, written

$$\mu \equiv \nu,$$

if both $\mu \ll \nu$ and $\nu \ll \mu$.

9.3.9. *Remarks.* In view of Definition 9.3.3 and the remarks following it, we have

$$\begin{aligned} \mu \equiv \nu &\Leftrightarrow \mu \ll |\nu| \text{ \& } \nu \ll |\mu| \\ &\Leftrightarrow |\mu| \ll |\nu| \text{ \& } |\nu| \ll |\mu| \Leftrightarrow |\mu| \equiv |\nu|; \end{aligned}$$

the latter condition says that the measures $|\mu|$ and $|\nu|$ have the same null sets, that is, for sets $E \in \mathcal{S}$,

$$|\mu|(E) = 0 \Leftrightarrow |\nu|(E) = 0,$$

equivalently

$$|\mu|_E = 0 \Leftrightarrow |\nu|_E = 0,$$

that is (9.1.23),

$$|\mu_E| = 0 \Leftrightarrow |\nu_E| = 0,$$

equivalently (9.1.23 again)

$$\mu_E = 0 \Leftrightarrow \nu_E = 0.$$

We approach the main decomposition theorem (Corollary 9.3.13) through three special cases (a lemma, a theorem and a corollary):

9.3.10. *Lemma.* If μ and ν are σ -finite measures on \mathcal{S} such that $\nu \ll \mu$, then there exists a set $A \in \mathcal{S}$ such that $\nu \equiv \mu_A$ and $\nu = \nu_A$.

Proof. By the Radon–Nikodym theorem, there exists a measurable function $g \geq 0$ such that $\nu = g \cdot \mu$ (see the proof of 9.2.15). Let

$$A = \{x : g(x) > 0\}.$$

Then $\varphi_A g = g$, therefore

$$\varphi_A \cdot \nu = \varphi_A \cdot (g \cdot \mu) = (\varphi_A g) \cdot \mu = g \cdot \mu = \nu$$

(the second equality by (8) of 9.2.4), thus $\nu = \varphi_A \cdot \nu = \nu_A$.

It will now suffice to show that $\nu \equiv \mu_A$. Indeed,

$$\begin{aligned}
 \nu(E) = 0 &\Leftrightarrow \nu_E = 0 \Leftrightarrow \varphi_E \cdot \nu = 0 \\
 &\Leftrightarrow \varphi_E \cdot (g \cdot \mu) = 0 \Leftrightarrow (\varphi_E g) \cdot \mu = 0 \\
 &\Leftrightarrow [(\varphi_E g) \cdot \mu](X) = 0 \quad (\text{cf. (1) of 9.2.4}) \\
 &\Leftrightarrow \int \varphi_E g d\mu = 0 \\
 &\Leftrightarrow \varphi_E g = 0 \quad \mu\text{-a.e.} \quad (4.4.21) \\
 &\Leftrightarrow \varphi_E \varphi_A = 0 \quad \mu\text{-a.e.} \Leftrightarrow \varphi_{E \cap A} = 0 \quad \mu\text{-a.e.} \\
 &\Leftrightarrow \mu(E \cap A) = 0 \Leftrightarrow \mu_A(E) = 0,
 \end{aligned}$$

thus ν and μ_A have the same null sets. \diamond

The Lemma is sufficient to yield the Lebesgue decomposition of a σ -finite measure ν with respect to a *finite* measure μ :

9.3.11. Theorem. *If μ and ν are measures on \mathcal{S} with μ finite and ν σ -finite, then there exist pairwise disjoint sets $A, B, C \in \mathcal{S}$ such that*

$$X = A \cup B \cup C, \quad \nu_A \equiv \mu_A, \quad \nu_B = 0, \quad \mu_C = 0,$$

whence $\nu = \nu_A + \nu_C$ is a (the!) Lebesgue decomposition of ν with respect to μ (unique by 9.3.7).

Proof. The measure $\mu + \nu$ is also σ -finite and $\nu \leq \mu + \nu$; obviously $\nu \ll \mu + \nu$, so by the Lemma there exists a set $E \in \mathcal{S}$ such that

$$(*) \quad \nu \equiv (\mu + \nu)_E, \quad \nu = \nu_E.$$

Writing $B = X - E$, we have $\nu_B = 0$ by (ii) of 9.3.2.

From (*) we see that

$$\nu \equiv (\mu + \nu)_E = \mu_E + \nu_E = \mu_E + \nu,$$

whence it is clear that $\mu_E \ll \nu$; applying the Lemma to the pair μ_E, ν , there exists a set $F \in \mathcal{S}$ such that

$$(**) \quad \mu_E \equiv \nu_F, \quad \mu_E = (\mu_E)_F = \mu_{E \cap F}.$$

Writing $A = E \cap F$ and citing (*) and (**) at the appropriate steps, we have

$$\nu_A = \nu_{E \cap F} = (\nu_E)_F = \nu_F \equiv \mu_E = \mu_{E \cap F} = \mu_A,$$

thus $\nu_A \equiv \mu_A$.

Since $A \subset E$ and $B = X - E$, we have $A \cap B = \emptyset$. Writing $C = E - A$, the measurable sets A, B, C are pairwise disjoint with $A \cup B \cup C = X$. Moreover, since $E = A \cup C$ and $\mu_A = \mu_E$ (noted in (**)), we have

$$\mu_E = \mu_{A \cup C} = \mu_A + \mu_C = \mu_E + \mu_C,$$

and $\mu_C = 0$ results on cancelling the finite measure μ_E .

Finally, we see from (*) and the disjointness of A and C that

$$\nu = \nu_E = \nu_{A \cup C} = \nu_A + \nu_C;$$

the relations $(\nu_C)_C = \nu_C$ and $\mu_C = 0$, that is,

$$(\nu_C)_C = \nu_C, \quad \mu_{X-C} = \mu,$$

show that $\nu_C \perp \mu$, whereas $\nu_A \equiv \mu_A \ll \mu$, thus $\nu = \nu_A + \nu_C$ is a Lebesgue decomposition of ν with respect to μ . \diamond

The extension to a pair of σ -finite measures is straightforward:

9.3.12. Corollary. (**Lebesgue decomposition**) *Same conclusion as the Theorem, assuming that both μ and ν are σ -finite measures.*

Proof. Let (E_n) be a sequence of pairwise disjoint sets in \mathcal{S} such that $X = \bigcup_{n=1}^{\infty} E_n$ and $\mu(E_n) < +\infty$ for all n . Then, for each n ,

$$\mu_{E_n} \text{ is finite, } \nu_{E_n} \text{ is } \sigma\text{-finite,}$$

so by the preceding theorem there exist pairwise disjoint sets U_n, V_n, W_n in \mathcal{S} with $X = U_n \cup V_n \cup W_n$, such that

$$(\nu_{E_n})_{U_n} \equiv (\mu_{E_n})_{U_n}, \quad (\nu_{E_n})_{V_n} = 0, \quad (\mu_{E_n})_{W_n} = 0,$$

that is,

$$(*) \quad \nu_{E_n \cap U_n} \equiv \mu_{E_n \cap U_n}, \quad \nu_{E_n \cap V_n} = 0, \quad \mu_{E_n \cap W_n} = 0.$$

For each n , the sets

$$A_n = E_n \cap U_n, \quad B_n = E_n \cap V_n, \quad C_n = E_n \cap W_n$$

are pairwise disjoint sets with

$$A_n \cup B_n \cup C_n = E_n,$$

therefore the sets

$$A = \bigcup_{n=1}^{\infty} A_n, \quad B = \bigcup_{n=1}^{\infty} B_n, \quad C = \bigcup_{n=1}^{\infty} C_n$$

are pairwise disjoint sets in \mathcal{S} with $A \cup B \cup C = \bigcup_{n=1}^{\infty} E_n = X$. By (*), we have

$$\nu_{A_n} \equiv \mu_{A_n}, \quad \nu_{B_n} = 0, \quad \mu_{C_n} = 0$$

for every n , whence it is clear from the countable additivity of measures that

$$\nu_A \equiv \mu_A, \quad \nu_B = 0, \quad \mu_C = 0,$$

thus the conclusions of the preceding theorem also hold for μ and ν . \diamond

Finally, the extension to signed measures is a triviality:

9.3.13. Corollary. *Same conclusion as the Theorem, assuming that μ and ν are signed measures such that the measures $|\mu|$ and $|\nu|$ are σ -finite.*

Proof. Apply the preceding corollary to the σ -finite measures $|\mu|$ and $|\nu|$: there exist pairwise disjoint sets A, B, C in \mathcal{S} with union X , such that

$$|\nu|_A \equiv |\mu|_A, \quad |\nu|_B = 0, \quad |\mu|_C = 0,$$

in other words (9.1.23)

$$|\nu_A| \equiv |\mu_A|, \quad |\nu_B| = 0, \quad |\mu_C| = 0,$$

that is,

$$|\nu_A| \equiv |\mu_A|, \quad \nu_B = 0, \quad \mu_C = 0,$$

and finally (9.3.9)

$$\nu_A \equiv \mu_A, \quad \nu_B = 0, \quad \mu_C = 0,$$

thus the conclusions of the Theorem also hold for μ and ν . \diamond

Exercises

1. Let λ be Lebesgue measure on the closed interval $[-1, 1]$, let $h : [-1, 1] \rightarrow \mathbb{R}$ be the function $h = -\varphi_{[-1, 0]} + \varphi_{(0, 1]}$ that is -1 on $[-1, 0]$ and $+1$ on $(0, 1]$, and let $\mu = h \cdot \lambda$ be the indefinite integral associated with the λ -integrable function h (4.7.1).

(i) $|\mu| = \lambda$, therefore $\lambda \ll \mu$ in the sense of Definition 9.3.3.

{Hint: 9.2.11.}

(ii) $\mu([-1, 1]) = 0$ but $\lambda([-1, 1]) = 2$.

2. If μ, ν are signed measures on \mathcal{S} such that the measures $|\mu|$ and $|\nu|$ are σ -finite, then the following conditions are equivalent:

(a) $\mu \equiv \nu$;

(b) $\nu = f \cdot |\mu|$ for some $f \in \mathcal{D}(|\mu|)$ such that $f(x) \neq 0$ $|\mu|$ -a.e.

{Hint: (a) \Rightarrow (b): Write $\nu = f \cdot |\mu|$ by the Radon-Nikodym theorem (9.2.15) and apply the equality $|\nu| = |f| \cdot |\mu|$ to the set $E = \{x : f(x) = 0\}$.

(b) \Rightarrow (a): For every $E \in \mathcal{S}$, $|\nu|(E) = \int \varphi_E |f| d|\mu|$; argue that $|\nu|(E) = 0 \Leftrightarrow \varphi_E = 0$ $|\mu|$ -a.e. (cf. 4.4.21).}

9.4. Convolution in $L^1(\mathbb{R})$

We know that, in the context of Lebesgue measure λ on \mathbb{R} (4.4.7), $L^1 = L^1_{\mathbb{C}}(\mathbb{R}, \mathcal{M}, \lambda)$ is a Banach space (6.4.18). In fact, there is a natural

way of introducing a product in L^1 in such a way that L^1 becomes a Banach algebra (in the sense of Definition 6.8.13) of capital importance in the theory of the Fourier transform.¹ The crux of the matter is this: given $u, v \in L^1$, say $u = \dot{f}$, $v = \dot{g}$ with $f, g \in \mathcal{L}^1 = \mathcal{L}^1_{\mathbb{C}}(\mathbb{R}, \mathcal{M}, \lambda)$ (see 6.4.14 for the dot notation), the problem is to create a function $h \in \mathcal{L}^1$ suitable for defining a product $uv = \dot{h}$. There is a *trivial* solution: always take $h = 0$; the result is a trivial algebra in which $uv = 0$ for all elements u and v . A useful solution must make h dependent on f and g in a subtler way: we shall show that there exists a function $h \in \mathcal{L}^1$ such that

$$(*) \quad h(x) = \int f(x-y)g(y)dy$$

for almost every x (here dy stands for Lebesgue measure λ in the context of y variable while x is fixed, and “almost every” refers to an exceptional set of Lebesgue measure zero). More precisely, we will show that the integral on the right side of $(*)$ exists for almost every x , and is equal to a Lebesgue-integrable function on the complement of a Lebesgue-measurable set (or of a Borel set) of measure zero; the proof is a deft application of the Fubini–Tonelli theorem (§7.4). For the proof that this leads to an associative (and commutative) algebra structure on L^1 , we refer the reader to more specialized texts.²

For the rest of the section, λ denotes Lebesgue measure, defined on the σ -algebra \mathcal{M} of Lebesgue-measurable sets (2.2.9); we write $\mu = \lambda|_{\mathcal{B}}$ for the restriction of λ to the σ -algebra $\mathcal{B} = \mathcal{B}(\mathbb{R})$ of Borel sets of \mathbb{R} (2.4.5). In the context of either λ or μ , we also use the notations $\int f(x, y)dx$ and $\int f(x, y)dy$ to indicate the variable of integration when the other variable in a function of two variables is held fixed.

Thus, there are two measure spaces in the picture: $(\mathbb{R}, \mathcal{M}, \lambda)$ and $(\mathbb{R}, \mathcal{B}, \mu)$. A complex-valued function $f : \mathbb{R} \rightarrow \mathbb{C}$ on \mathbb{R} is called *Lebesgue-measurable* if it is measurable with respect to \mathcal{M} , and *Borel* if it is measurable with respect to \mathcal{B} (cf. 9.4.3 below). On the one hand, every μ -integrable (Borel) function $f : \mathbb{R} \rightarrow \mathbb{C}$ is obviously λ -integrable (i.e., Lebesgue-integrable), with $\int f d\lambda = \int f d\mu$. On the other hand, if $f : \mathbb{R} \rightarrow \mathbb{C}$ is λ -integrable then there exists a Borel function $g : \mathbb{R} \rightarrow \mathbb{C}$ such that $f = g$ on the complement of a Borel set of measure zero (see Proposition 9.4.7 below); in particular, $f = g$ λ -a.e. (in fact μ -a.e.) and g is μ -integrable with $\int g d\mu = \int f d\lambda$. Thus, although $\mathcal{L}^1_{\mathbb{C}}(\mu)$ is a proper linear subspace of $\mathcal{L}^1_{\mathbb{C}}(\lambda)$, the Banach spaces $L^1_{\mathbb{C}}(\mu)$ and $L^1_{\mathbb{C}}(\lambda)$ may be identified. For technical reasons explained below, it will be convenient to

¹ Cf. E. Hewitt and K. Stromberg, *Real and abstract analysis* [Springer, New York, 1965], p. 250, (16.36).

² Cf. L.H. Loomis, *An introduction to abstract harmonic analysis* [Van Nostrand, New York, 1953], E. Hewitt and K. Stromberg, *op. cit.*, p. 399, (21.34), or the author, *Measure and integration* [Macmillan, New York, 1965; reprinted Chelsea, New York, 1970], §86.

discuss Borel functions first, then adapt the results to Lebesgue-measurable functions.

Some preparatory material on Borel functions is in order. Recall that if X is any topological space, the class $\mathcal{B}(X)$ of *Borel sets* of X is the σ -algebra generated by the open sets of X (4.1.2,(ii)). In particular, we abbreviate $\mathcal{B} = \mathcal{B}(\mathbb{R})$ for the σ -algebra of Borel sets of \mathbb{R} (cf. 2.4.11). A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is called a *Borel function* if $f^{-1}(\mathcal{B}) \subset \mathcal{B}$ (4.1.10); in other words f , regarded as a function defined on the measurable space $(\mathbb{R}, \mathcal{B})$, is measurable with respect to \mathcal{B} (4.1.3)—equivalently (4.1.6),

$$U \text{ open in } \mathbb{R} \Rightarrow f^{-1}(U) \in \mathcal{B}.$$

This suggests the following generalization:

9.4.1. *Definition.* Let X and Y be topological spaces. A function $f : X \rightarrow Y$ is said to be a **Borel function** if $f^{-1}(\mathcal{B}(Y)) \subset \mathcal{B}(X)$, that is,

$$B \text{ Borel in } Y \Rightarrow f^{-1}(B) \text{ Borel in } X;$$

equivalently (cf. the proof of 4.1.6), $f^{-1}(U)$ is a Borel set in X for every open set U in Y .

9.4.2. *Remarks.* (i) Every continuous function is a Borel function. {The inverse image of an open set is open, hence is a Borel set.}

(ii) If $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ are Borel functions (X, Y, Z topological spaces), then the composite function $g \circ f : X \rightarrow Z$ is Borel. {If B is a Borel set in Z then $(g \circ f)^{-1}(B) = f^{-1}(g^{-1}(B))$ is the inverse image under f of a Borel set in Y .}

(iii) If X and Y are topological spaces, then

$$\mathcal{B}(X) \times \mathcal{B}(Y) \subset \mathcal{B}(X \times Y).$$

{If U, V are open sets in X, Y , respectively, then $U \times V$ is open in $X \times Y$ hence is a Borel set.}

(iv) If X and Y are topological spaces having a countable base for the topology (6.1.20)—for example if X and Y are separable metric spaces (6.1.21)—then

$$\mathcal{B}(X) \times \mathcal{B}(Y) = \mathcal{B}(X \times Y).$$

{Every open set in $X \times Y$ is the union of a sequence of sets $U_n \times V_n$ with U_n, V_n open in X, Y , respectively, whence $\mathcal{B}(X \times Y) \subset \mathcal{B}(X) \times \mathcal{B}(Y)$.} In particular,

$$\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R} \times \mathbb{R}).$$

We are particularly interested in Borel functions in the case that $Y = \mathbb{C}$ (and, ultimately, $X = \mathbb{R}$):

9.4.3. *Proposition.* Let X be a topological space. The following conditions on a function $f : X \rightarrow \mathbb{C}$ are equivalent:

- (a) f is a Borel function (in the sense of Definition 9.4.1);
 (b) f is measurable with respect to the σ -algebra $\mathcal{B}(X)$ (in the sense of Definition 6.4.1);
 (c) the functions $\operatorname{Re} f, \operatorname{Im} f : X \rightarrow \mathbb{R}$ are measurable with respect to $\mathcal{B}(X)$ (in the sense of Definition 4.1.3);
 (d) the functions $\operatorname{Re} f, \operatorname{Im} f : X \rightarrow \mathbb{R}$ are Borel (in the sense of 9.4.1).

Proof. (b) \Leftrightarrow (c) by the Definition of (b) (6.4.1).

(c) \Leftrightarrow (d) by the remarks in the paragraph preceding 9.4.1.

Recall that \mathbb{C} and \mathbb{R}^2 may be identified as metric spaces (cf. 3.1.9, 3.1.15), hence as topological spaces.

(a) \Rightarrow (d): If U is an open set in \mathbb{R} , then the ‘rectangle’ $U + i\mathbb{R}$ (in the Gaussian plane $\mathbb{C} = \mathbb{R}^2$) is open in \mathbb{C} , so by the hypothesis (a), $f^{-1}(U + i\mathbb{R})$ is a Borel set in X ; since

$$f^{-1}(U + i\mathbb{R}) = (\operatorname{Re} f)^{-1}(U) \cap (\operatorname{Im} f)^{-1}(\mathbb{R}) = (\operatorname{Re} f)^{-1}(U),$$

we see that $(\operatorname{Re} f)^{-1}(U) \in \mathcal{B}(X)$ for all open sets U in \mathbb{R} , thus $\operatorname{Re} f$ is a Borel function (in the sense of 9.4.1). Similarly $\operatorname{Im} f$ is a Borel function (consider $\mathbb{R} + iU$).

(d) \Rightarrow (a): Every open set W in \mathbb{C} is the union of open rectangles with Gaussian-rational vertices, hence can be expressed as a countable union

$$W = \bigcup_{n=1}^{\infty} (U_n + iV_n),$$

where U_n and V_n are open intervals in \mathbb{R} ; by the hypothesis (d), the set

$$f^{-1}(W) = \bigcup_{n=1}^{\infty} f^{-1}(U_n + iV_n) = \bigcup_{n=1}^{\infty} (\operatorname{Re} f)^{-1}(U_n) \cap (\operatorname{Im} f)^{-1}(V_n)$$

is clearly a Borel set in X . Thus f is a Borel function (in the sense of 9.4.1). \diamond

A bonus of criterion (b) is that the set of all Borel functions $X \rightarrow \mathbb{C}$ is a complex algebra for the pointwise operations:

9.4.4. Corollary. Let X be a topological space. If $f, g : X \rightarrow \mathbb{C}$ are Borel functions, $c \in \mathbb{R}$ and $\alpha > 0$, then the functions $f + g$, cf , fg and $|f|^\alpha$ are also Borel functions.

Proof. This is immediate from 6.4.2 and criterion (b) of Proposition 9.4.3. \diamond

9.4.5. Corollary. If X is a topological space and $f_n : X \rightarrow \mathbb{C}$ is a pointwise convergent sequence of Borel functions, then the limit function $\lim f_n$ is also a Borel function.

Proof. Immediate from 6.4.3 and criterion (b) of 9.4.3. \diamond

9.4.6. Corollary. If X and Y are topological spaces and $f : X \rightarrow \mathbb{C}$, $g : Y \rightarrow \mathbb{C}$ are Borel functions, then the function $f \otimes g : X \times Y \rightarrow \mathbb{C}$ defined by

$$(f \otimes g)(x, y) = f(x)g(y) \quad (x \in X, y \in Y)$$

is measurable with respect to $\mathcal{B}(X) \times \mathcal{B}(Y)$; in particular, $f \otimes g$ is a Borel function.

Proof. By criterion (b) of 9.4.3, our hypothesis is that f, g are measurable with respect to $\mathcal{B}(X), \mathcal{B}(Y)$, respectively. If f and g are characteristic functions, say $f = \varphi_E, g = \varphi_F$ with E, F Borel sets in X, Y , respectively, then $f \otimes g = \varphi_{E \times F}$ is measurable with respect to $\mathcal{B}(X) \times \mathcal{B}(Y)$ because $E \times F \in \mathcal{B}(X) \times \mathcal{B}(Y)$. The case that f and g are simple (with respect to $\mathcal{B}(X)$ and $\mathcal{B}(Y)$, respectively) then follows from the bilinearity of the operation \otimes . In the general case, f and g are the pointwise limits of sequences f_n, g_n of simple Borel functions (because their real and imaginary parts are, by criterion (b) of 9.4.3 and Theorem 4.1.26), therefore

$$f \otimes g = \lim f_n \otimes g_n$$

is measurable with respect to $\mathcal{B}(X) \times \mathcal{B}(Y)$ by Proposition 6.4.3.

Finally, $(f \otimes g)^{-1}(\mathcal{B}(\mathbb{C})) \subset \mathcal{B}(X) \times \mathcal{B}(Y) \subset \mathcal{B}(X \times Y)$ by (iii) of 9.4.2, thus $f \otimes g$ is a Borel function on $X \times Y$. \diamond

Our last generality before we get down to business (convolution):

9.4.7. Proposition. If $f : \mathbb{R} \rightarrow \mathbb{C}$ is Lebesgue-measurable, then there exists a Borel function $g : \mathbb{R} \rightarrow \mathbb{C}$ such that $f = g$ on the complement of a Borel set of measure 0 (thus $f = g$ μ -a.e. and hence λ -a.e.).

Proof. The assumption on f is that its real and imaginary parts are measurable with respect to the σ -algebra \mathcal{M} of Lebesgue-measurable subsets of \mathbb{R} (4.1.5, (iii) and 6.4.1). In view of Proposition 9.4.3, we can suppose that $f : \mathbb{R} \rightarrow \mathbb{R}$, and by Theorem 4.1.15 we can suppose further that $f \geq 0$. Contemplating Theorem 4.1.26 and Corollary 4.1.20, we can suppose in addition that f is simple; finally, taking into account Theorem 4.1.9, we can suppose that f is the characteristic function of a Lebesgue-measurable set. (At every step of the reduction, we note that a union of a finite or denumerable set of Borel sets of measure 0 is also such a set.) Say $f = \varphi_E, E \in \mathcal{M}$. By Corollary 2.4.15, there exist Borel sets F and G such that $F \subset E \subset G$ and $\lambda(G - F) = 0$. The function $g = \varphi_F$ is Borel, $f = g = 1$ on F and $f = g = 0$ on $\mathbb{C}G$, thus $f = g$ on the complement $\mathbb{C}(G - F) = F \cup \mathbb{C}G$ of the Borel set $G - F$ of measure 0. (The argument shows that every Lebesgue-measurable set of measure 0

is contained in a Borel set of measure 0, thus the concepts “ λ -a.e.” and “ μ -a.e.” coincide—and may indifferently be written “a.e.”) \diamond

9.4.8. *Definition.* If $f, g : \mathbb{R} \rightarrow \mathbb{C}$ we define a function $f \nabla g : \mathbb{R}^2 \rightarrow \mathbb{C}$ by the formula

$$(f \nabla g)(x, y) = f(x - y)g(y) \quad (x, y \in \mathbb{R})$$

(note that the arguments of f and g on the right side add up to x); writing $\pi : \mathbb{R}^2 \rightarrow \mathbb{R}$ for the function $\pi(x, y) = x - y$, we have

$$(f \nabla g)(x, y) = f(\pi(x, y))g(y) = (f \circ \pi)(x, y) \cdot (1 \otimes g)(x, y),$$

thus $f \nabla g = (f \circ \pi)(1 \otimes g)$.

9.4.9. *Proposition.* If $f, g : \mathbb{R} \rightarrow \mathbb{C}$ are Borel functions, then the function $f \nabla g : \mathbb{R}^2 \rightarrow \mathbb{C}$ defined above is measurable with respect to $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$, thus is a Borel function on \mathbb{R}^2 .

Proof. Since $\pi : \mathbb{R}^2 \rightarrow \mathbb{R}$ is continuous, $f \circ \pi : \mathbb{R}^2 \rightarrow \mathbb{C}$ is a Borel function on \mathbb{R}^2 by (i) and (ii) of 9.4.2; in other words, by (iv) of 9.4.2, $f \circ \pi$ is measurable with respect to $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$. (This is the step where Borel sets are more convenient than Lebesgue-measurable sets.) On the other hand, $1 \otimes g$ is measurable with respect to $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$ by Corollary 9.4.6, therefore so is the pointwise product $(f \circ \pi)(1 \otimes g) = f \nabla g$. \diamond

Recall that $\mu = \lambda|_{\mathcal{B}}$ denotes the restriction of Lebesgue measure λ to the σ -algebra of Borel sets $\mathcal{B} = \mathcal{B}(\mathbb{R})$.

Prior to calculating the iterated integrals of $f \nabla g$ (with respect to $\mu \times \mu$), we describe the sections (cf. 7.3.1) of $f \nabla g$:

9.4.10. *Lemma.* If $f, g : \mathbb{R} \rightarrow \mathbb{C}$ then, for all $x, y \in \mathbb{R}$,

$$(f \nabla g)_x = (f \circ \pi)_x \cdot g, \quad (f \nabla g)^y = g(y)f_{-y},$$

where f_{-y} denotes the translate of f by $-y$, that is, $f_{-y}(x) = f(x - y)$ for all $x \in \mathbb{R}$.

Proof. For all $x, y \in \mathbb{R}$,

$$\begin{aligned} (f \nabla g)_x(y) &= (f \nabla g)(x, y) = (f \circ \pi)(x, y)g(y) = (f \circ \pi)_x(y) \cdot g(y), \\ (f \nabla g)^y(x) &= (f \nabla g)(x, y) = f(x - y)g(y) = g(y)f_{-y}(x) \end{aligned}$$

by the formulas of Definition 9.4.8. \diamond

9.4.11. *Lemma.* Let $f, g : \mathbb{R} \rightarrow \mathbb{C}$.

(i) If $f \in \mathcal{L}^1_{\mathbb{C}}$ (for either λ or μ) then, for every $y \in \mathbb{R}$, $(f \nabla g)^y \in \mathcal{L}^1_{\mathbb{C}}$ and

$$\int (f \nabla g)^y = \int f(x - y)g(y)dx = \left(\int f \right) g(y).$$

(ii) If $f, g \in \mathcal{L}_{\mathbb{C}}^2$ (for either λ or μ) then, for every $x \in \mathbb{R}$,

$$(f \circ \pi)_x \in \mathcal{L}_{\mathbb{C}}^2, \quad (f \nabla g)_x = (f \circ \pi)_x \cdot g \in \mathcal{L}_{\mathbb{C}}^1$$

and

$$\int f(x-y)g(y)dy = \int (f \circ \pi)_x(y)g(y)dy.$$

Proof. (i) By the second formula of the preceding lemma, $(f \nabla g)^y = g(y)f_{-y}$ for every $y \in \mathbb{R}$. From the translation-invariance of Lebesgue measure (2.1.10) it is straightforward to show that $f_{-y} \in \mathcal{L}_{\mathbb{C}}^1$ and $\int f_{-y} = \int f$ (first consider the case that f is a characteristic function, then a simple function ≥ 0 , etc.).

(ii) Let $x \in \mathbb{R}$. By the first formula of the preceding lemma, $(f \nabla g)_x = (f \circ \pi)_x \cdot g$; since

$$(f \circ \pi)_x(y) = (f \circ \pi)(x, y) = f(x-y)$$

and the function $|f|^2$ is by assumption integrable (6.7.1), it is straightforward to show, using the invariance of Lebesgue measure under the transformations $y \mapsto -y \mapsto x-y$ (2.1.11, 2.1.10), that $(f \circ \pi)_x \in \mathcal{L}_{\mathbb{C}}^2$. It follows that $(f \nabla g)_x = (f \circ \pi)_x \cdot g$ is integrable by Hölder's inequality (6.7.2) and

$$\int f(x-y)g(y)dy = \int (f \nabla g)_x(y)dy = \int (f \circ \pi)_x(y)g(y)dy.$$

{Incidentally, the right member of the equality can be written as the inner product in $L_{\mathbb{C}}^2$ of the equivalence classes of $(f \circ \pi)_x$ and \bar{g} (see §6.7, Exercise 5).} \diamond

9.4.12. Theorem. If $f, g \in \mathcal{L}_{\mathbb{C}}^1(\mu)$ then $f \nabla g \in \mathcal{L}_{\mathbb{C}}^1(\mu \times \mu)$ and

$$\int (f \nabla g)d(\mu \times \mu) = \left(\int f d\mu \right) \left(\int g d\mu \right).$$

Proof. Since the operation $f \nabla g$ (hence its integral) is bilinear, as is the expression on the right side, we can suppose that $f \geq 0$ and $g \geq 0$ (express the real and imaginary parts of f, g as a difference of positive integrable Borel functions); then also $f \nabla g \geq 0$. From 9.4.9 we know that $f \nabla g$ is measurable with respect to the domain $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$ of $\mu \times \mu$. By (i) of the preceding lemma, $(f \nabla g)^y$ is μ -integrable for every $y \in \mathbb{R}$, and

$$\int (f \nabla g)^y d\mu = \left(\int f d\mu \right) g(y).$$

The function $\left(\int f d\mu \right) g$ being μ -integrable, we see (Definition 7.3.3) that the iterated integral

$$\iint (f \nabla g)(x, y) dx dy$$

exists and is equal to

$$\int \left(\int f d\mu \right) g d\mu = \left(\int f d\mu \right) \left(\int g d\mu \right).$$

By the Fubini–Tonelli theorem (7.4.7), $f \nabla g$ is $(\mu \times \mu)$ -integrable with integral $(\int f d\mu)(\int g d\mu)$. \diamond

9.4.13. Corollary. If $f, g \in \mathcal{L}_\mathbb{C}^1(\mu)$ then the iterated integral

$$\iint (f \nabla g)(x, y) dy dx$$

exists and is equal to $(\int f d\mu)(\int g d\mu)$.

Proof. This is immediate from the preceding theorem and Fubini's theorem (7.4.8). \diamond

9.4.14. Remarks. The preceding corollary can be expressed as follows: If f and g are Lebesgue-integrable Borel functions, then there exist a Lebesgue-integrable Borel function h and a Borel set E of measure 0 such that

$$x \in \mathbb{R} - E \Rightarrow \exists \int f(x - y)g(y)dy = h(x)$$

and $\int h d\lambda = (\int f d\lambda)(\int g d\lambda)$.

9.4.15. Definition. Let $f, g : \mathbb{R} \rightarrow \mathbb{C}$. The **convolution** of f and g (in that order—but see Exercise 2) is the function $f * g$ whose domain is the set

$$\begin{aligned} D_{f*g} &= \{x \in \mathbb{R} : (f \nabla g)_x \in \mathcal{L}_\mathbb{C}^1(\lambda)\} \\ &= \{x \in \mathbb{R} : \text{the function } y \mapsto f(x - y)g(y) \text{ is } \lambda\text{-integrable}\} \end{aligned}$$

and whose values are given by the formula

$$(f * g)(x) = \int f(x - y)g(y)dy \quad (x \in D_{f*g}).$$

9.4.16. Remarks. The message of 9.4.14: If f and g are Lebesgue-integrable Borel functions, then there exist a Lebesgue-integrable Borel function h and a Borel set E of measure 0 such that

$$D_{f*g} \supset \mathbb{R} - E \quad \text{and} \quad f * g = h \quad \text{on } \mathbb{R} - E.$$

{It can be shown³ that D_{f*g} is itself a Borel set (whose complement has measure 0).} Moreover, $\int h d\lambda = (\int f d\lambda)(\int g d\lambda)$. Abusing the notations slightly, $\int (f * g) d\lambda = (\int f d\lambda)(\int g d\lambda)$.

³ Cf. the author, *op. cit.*, p. 293, Theorem 1.

Convolution is an operation that ‘smoothes away’ behavior on negligible sets:

9.4.17. Lemma. *If $f_1, f_2, g_1, g_2 : \mathbb{R} \rightarrow \mathbb{C}$ are complex-valued functions on \mathbb{R} such that*

$$f_1 = f_2 \text{ a.e.} \quad \text{and} \quad g_1 = g_2 \text{ a.e.}$$

*(with respect to Lebesgue measure) then $f_1 * g_1 = f_2 * g_2$ (including equality of the domains).*

Proof. For each $x \in \mathbb{R}$, the functions

$$y \mapsto f_1(x - y)g_1(y), \quad y \mapsto f_2(x - y)g_2(y)$$

are equal a.e.; since Lebesgue measure is complete (§2.4, Exercise 8), each of these functions is Lebesgue-integrable (or Lebesgue-measurable) if and only if the other is, and their integrals are then equal. \diamond

The extension of convolution to Lebesgue-integrable (not necessarily Borel) functions is now effortless:

9.4.18. Theorem. *If $f, g : \mathbb{R} \rightarrow \mathbb{C}$ are Lebesgue-integrable, then there exist a Lebesgue-integrable Borel function h and a Borel set E of measure 0 such that*

$$D_{f*g} \supset \mathbb{R} - E \quad \text{and} \quad f * g = h \text{ on } \mathbb{R} - E.$$

Proof. Let $f_0, g_0 : \mathbb{R} \rightarrow \mathbb{C}$ be Borel functions such that $f = f_0$ and $g = g_0$ on the complement of a Borel set of measure 0 (9.4.7). Then f_0, g_0 are also Lebesgue-integrable and, as noted in 9.4.16, there exist an integrable Borel function h and a Borel set E of measure 0 such that

$$D_{f_0*g_0} \supset \mathbb{R} - E \quad \text{and} \quad f_0 * g_0 = h \text{ on } \mathbb{R} - E.$$

By the lemma, $f * g = f_0 * g_0$, thus the function h and the Borel set E of measure 0 meet the requirements of the theorem. \diamond

9.4.19. Remark. With notations as in the theorem, we have shown that the iterated integral

$$\iint (f \nabla g)(x, y) dy dx$$

exists (note that in Definition 7.3.3, the function of two variables is not required to be measurable), without knowing whether $f \nabla g$ is measurable with respect to $\mathcal{M} \times \mathcal{M}$. The Borel functions f_0, g_0 in the foregoing proof have done the heavy lifting (via 9.4.12 and 9.4.13).

Exercises

1. There exist compact spaces X and Y for which the inclusion $\mathcal{B}(X) \times \mathcal{B}(Y) \subset \mathcal{B}(X \times Y)$ of 9.4.2, (iii) is proper.⁴

2. (i) (*Commutative law*) If $f, g: \mathbb{R} \rightarrow \mathbb{C}$ are Lebesgue-integrable, then $f * g = g * f$ a.e.⁵

(ii) (*Associative law*) If $f, g, h: \mathbb{R} \rightarrow \mathbb{C}$ are Lebesgue-integrable, and if $p, q: \mathbb{R} \rightarrow \mathbb{C}$ are Lebesgue-integrable functions such that $f * g = p$ a.e. and $g * h = q$ a.e., then $p * h = f * q$ a.e.⁶

3. With notations as in Lemma 9.4.11, if (with respect to either λ or μ) one of f, g is in $\mathcal{L}_\mathbb{C}^1$ and the other is in $\mathcal{L}_\mathbb{C}^\infty$, then $(f \nabla g)_x \in \mathcal{L}_\mathbb{C}^1$ for all $x \in \mathbb{R}$.

{Hint: Lemma 6.6.8 and the formula $(f \nabla g)_x = (f \circ \pi)_x \cdot g \cdot$ }

4. If $u, v \in L_\mathbb{C}^1 = L_\mathbb{C}^1(\mathbb{R}, \mathcal{M}, \lambda)$, write $u = \dot{f}$, $v = \dot{g}$ as in 6.4.14, choose $h \in \mathcal{L}_\mathbb{C}^1$ so that $h = f * g$ a.e. (9.4.18), and define a product in $L_\mathbb{C}^1$ by the formula $uv = h$. Then the function $\varphi: L_\mathbb{C}^1 \rightarrow \mathbb{C}$ defined by $\varphi(u) = \int f \, d\lambda$ defines a linear form on $L_\mathbb{C}^1$ such that $\varphi(uv) = \varphi(u)\varphi(v)$ for all u, v . Moreover, $|\varphi(u)| \leq \|u\|_1$, consequently φ is continuous.

9.5. Integral Operators (with Continuous Kernel Function)

This section is an application of Ascoli's theorem (§8.1) to certain linear mappings in function spaces derived from Lebesgue measure. As these spaces are equipped with a seminorm (rather than a norm), some generalities on such matters are in order.

Recall that a *seminorm* on a real or complex vector space E is a function $x \mapsto \|x\|$ on E satisfying all conditions of a norm except that $\|x\|$ may be 0 for a nonzero vector x (6.4.9); thus, $\|x + y\| \leq \|x\| + \|y\|$ (*subadditivity*), $\|cx\| = |c| \|x\|$ for all scalars c and all vectors x (*absolute homogeneity*) and $\|x\| \geq 0$ for all vectors x (*nonnegativity*). The pair $(E, \|\cdot\|)$ is called a *seminormed space*.

9.5.1. *Example.* (i) Every normed space is a seminormed space.

(ii) In the vector space \mathbb{C}^n of all n 'ples $x = (x_1, \dots, x_n)$ of complex numbers, for each index k the function $x \mapsto |x_k|$ is a seminorm.

(iii) If $1 \leq p \leq +\infty$, the vector spaces $\mathcal{L}_\mathbb{R}^p$ and $\mathcal{L}_\mathbb{C}^p$ relative to a measure space (X, \mathcal{S}, μ) are seminormed spaces for the functionals¹

⁴Cf. the author, *op. cit.*, p. 183, Exercise 16; H.L. Royden, *Real analysis* [3rd edn., Macmillan, New York, 1988], p. 336, Exercise 7.

⁵See E. Hewitt and K. Stromberg, *op. cit.*, p. 397, (21.32).

⁶Cf. the author, *op. cit.*, p. 283, Theorem 2.

¹The term 'functional' is commonly used for a scalar-valued function defined on a set of functions.

$f \mapsto \|f\|_p$ defined in 6.4.7, 6.6.3 or 6.7.1 (according as $p = 1$, $p = +\infty$ or $1 < p < +\infty$). For $f \in \mathcal{L}^p$, $\|f\|_p = 0 \Leftrightarrow f = 0$ μ -a.e. (4.4.21, 6.6.5), thus $f \mapsto \|f\|_p$ is a norm (that is, $f = 0$ μ -a.e. $\Leftrightarrow f = 0$) if and only if the empty set is the only set in \mathcal{S} of measure 0.

9.5.2. *Definition.* Let $(E, \|\cdot\|)$ be a seminormed space. A sequence (x_n) in E is said to be **Cauchy** if

$$\|x_m - x_n\| \rightarrow 0 \text{ in } \mathbb{R} \text{ as } m, n \rightarrow \infty,$$

and it is said to be **convergent** if there exists a vector $x \in E$ such that

$$\|x_n - x\| \rightarrow 0 \text{ in } \mathbb{R} \text{ as } n \rightarrow \infty,$$

in which case x_n is said to **converge** to the **limit** x , written $x_n \rightarrow x$. If every Cauchy sequence in E is convergent, then E is said to be a **complete** seminormed space (or, when $\|\cdot\|$ is a norm, a **Banach space**; cf. 6.4.17).

9.5.3. *Remarks.* (i) In a seminormed space, every convergent sequence is Cauchy.

(ii) Limits in a seminormed space $(E, \|\cdot\|)$ need not be unique: if $x_n \rightarrow x$ in E and if $y \in E$, then $x_n \rightarrow y \Leftrightarrow \|x - y\| = 0$.

(iii) If $x_n \rightarrow x$ in a seminormed space $(E, \|\cdot\|)$, then $\|x_n\| \rightarrow \|x\|$ (in particular, the sequence $\|x_n\|$ is bounded in \mathbb{R}).

(iv) If (x_n) is a Cauchy sequence in a seminormed space $(E, \|\cdot\|)$ then the sequence $\|x_n\|$ in \mathbb{R} is Cauchy (hence convergent).

(v) The spaces \mathcal{L}^p of 9.5.1, (iii) are complete (6.4.13, 6.6.7, 6.7.4).

{Proofs: (iv) For all vectors x, y , $|\|x\| - \|y\|| \leq \|x - y\|$; for,

$$\|x\| = \|(x - y) + y\| \leq \|x - y\| + \|y\|$$

and similarly $\|y\| \leq \|y - x\| + \|x\|$. Thus if $\|x_m - x_n\| \rightarrow 0$ as $m, n \rightarrow \infty$, then also $|\|x_m\| - \|x_n\|| \rightarrow 0$ as $m, n \rightarrow \infty$.

(iii) $|\|x_n\| - \|x\|| \leq \|x_n - x\|$.

(i) If $\|x_n - x\| \rightarrow 0$ then $\|x_m - x_n\| \leq \|x_m - x\| + \|x - x_n\| \rightarrow 0$ as $m, n \rightarrow \infty$.

(ii) $\Rightarrow: \|x - y\| \leq \|x - x_n\| + \|x_n - y\| \rightarrow 0$.

$\Leftarrow: \|x_n - y\| \leq \|x_n - x\| + \|x - y\| = \|x_n - x\| \rightarrow 0$.}

When $1 \leq p < +\infty$, convergence in \mathcal{L}^p is called *convergence in mean* (of order p), and Cauchy sequences are said to be *Cauchy in mean* (of order p) (cf. 6.4.13).

9.5.4. *Definition.* Let $(E, \|\cdot\|)$ and $(F, \|\cdot\|)$ be seminormed spaces and let $T: E \rightarrow F$ be a linear mapping; T is said to be **continuous** if

$$x_n \rightarrow x \text{ in } E \Rightarrow Tx_n \rightarrow Tx \text{ in } F,$$

and **compact** if

$$\|x_n\| \text{ bounded} \Rightarrow Tx_n \text{ has a convergent subsequence.}$$

(The reason for restricting the definitions to linear mappings will be clear from what follows, including the exercises—for example, Exercise 2.)

9.5.5. Proposition. With E and F as in the preceding definition, let $T : E \rightarrow F$ be a linear mapping and let $B = \{x \in E : \|x\| \leq 1\}$ (called the 'closed unit ball' of E). Then:

(1) T is continuous $\Leftrightarrow T(B)$ is bounded, in the sense that $\{\|Tx\| : x \in B\}$ is a bounded subset of \mathbb{R} .

(2) T is compact \Leftrightarrow every sequence in $T(B)$ has a convergent subsequence.

(3) T compact $\Rightarrow T$ continuous.

(4) If T is continuous and $x \in E$, then $\|x\| = 0 \Rightarrow \|Tx\| = 0$.

Proof. (1) \Rightarrow : Arguing contrapositively, suppose $T(B)$ is unbounded. Choose a sequence (x_n) in B such that $\|Tx_n\| \geq n$ for all n . Then $\|\frac{1}{n}x_n\| = \frac{1}{n}\|x_n\| \leq 1/n \rightarrow 0$, so $\frac{1}{n}x_n \rightarrow 0$; but $\|T(\frac{1}{n}x_n)\| = \frac{1}{n}\|Tx_n\| \geq 1$ for all n , thus $T(\frac{1}{n}x_n) \not\rightarrow 0 = T0$, consequently T is not continuous.

\Leftarrow : By assumption, there exists a constant $M > 0$ such that $\|Tx\| \leq M$ for all $x \in B$. Suppose $x_n \rightarrow x$ in E , that is, $\|x_n - x\| \rightarrow 0$. Given any $\epsilon > 0$, we know that $\|x_n - x\| \leq \epsilon/M$ ultimately, therefore $\|(M/\epsilon)(x_n - x)\| \leq 1$ ultimately, that is, $(M/\epsilon)(x_n - x) \in B$ ultimately; by the definition of M , $\|T[(M/\epsilon)(x_n - x)]\| \leq M$ ultimately, whence $\|Tx_n - Tx\| \leq \epsilon$ ultimately (by the linearity of T and the absolute homogeneity of the seminorm), thus $Tx_n \rightarrow Tx$.

(2) \Rightarrow : Immediate from Definition 9.5.4.

\Leftarrow : Suppose (x_n) is a sequence in E that is bounded, say $\|x_n\| \leq M$ for all n , where $M > 0$. Then $\|M^{-1}x_n\| \leq 1$ for all n , so by assumption $T(M^{-1}x_n)$ has a convergent subsequence, say $T(M^{-1}x_{n_k}) \rightarrow y$ in F . Then $Tx_{n_k} \rightarrow My$, as one sees from the computation

$$\|Tx_{n_k} - My\| = \|M[T(M^{-1}x_{n_k}) - y]\| = M\|T(M^{-1}x_{n_k}) - y\| \rightarrow M \cdot 0 = 0.$$

(3) Arguing contrapositively, suppose T is not continuous. By (1), there exists a sequence $x_n \in B$ such that $\|Tx_n\| \geq n$ for all n . Although $\|x_n\| \leq 1$ for all n , the sequence Tx_n can have no convergent subsequence; for, $Tx_{n_k} \rightarrow y$ would imply that $\|Tx_{n_k}\| \rightarrow \|y\|$, contrary to $\|Tx_{n_k}\| \geq n_k$. Thus T is not compact.

(4) Assuming T continuous, suppose $\|x\| = 0$. Then $\|nx\| = n\|x\| = 0 \leq 1$ shows that $nx \in B$ for all n , therefore $n\|Tx\| = \|T(nx)\|$ is bounded by (1), whence $\|Tx\| = 0$. \diamond

The preliminaries are over; let's get down to business.

Notations fixed for the rest of the section:

$I = [a, b]$ is a closed interval in \mathbb{R} , with $a < b$.

$S = \mathcal{M} \cap I$ is the σ -algebra of all Lebesgue-measurable subsets of I (4.7.6); the restriction $\lambda|_S$ of Lebesgue measure λ to S will also be denoted λ (abuse of notation!) and, instead of $d\lambda$, we may write dx to

indicate integration with respect to a particular variable x in a function of several variables.

Fix a number p , $1 \leq p \leq +\infty$. In the context of the measure space $(I, \mathcal{S}, \lambda)$ just described, we write $\mathcal{L}^p = \mathcal{L}_{\mathbb{C}}^p(I, \mathcal{S}, \lambda)$ for the seminormed space cited in 9.5.1, (iii) and 9.5.3, (v).

We write $\mathcal{C} = \mathcal{C}_{\mathbb{C}}(I)$ for the set of all continuous complex-valued functions $f : I \rightarrow \mathbb{C}$, equipped with the sup-norm $f \mapsto \|f\|_{\infty}$ (6.8.12).

9.5.6. *Remarks.* (i) Equipped with the pointwise operations and the sup-norm $f \mapsto \|f\|_{\infty}$, \mathcal{C} is a complex Banach algebra (6.8.14).

(ii) \mathcal{C} is a linear subspace of \mathcal{L}^p . For $p = +\infty$ this is immediate from the Lebesgue- (even Borel-) measurability of a continuous function on I ; moreover, the notation $\|f\|_{\infty}$ represents the same number in the context of either \mathcal{C} or \mathcal{L}^{∞} since, for a continuous function f , $|f| \leq M$ μ -a.e. is equivalent to $|f| \leq M$ (if $|f| \leq M$ μ -a.e. then the set of points where $|f(x)| > M$ is a negligible open set in I , hence is empty). On the other hand, suppose $1 \leq p < +\infty$; a function $f \in \mathcal{C}$ is a bounded measurable function on a finite measure space, hence is integrable, and this observation applied to $|f|^p$ shows that $f \in \mathcal{L}^p$.

(iii) Suppose $f \in \mathcal{L}^p$ and $g \in \mathcal{L}^q$, where q is the exponent complementary to p . {Convention: when $p = 1$, $q = +\infty$; when $p = +\infty$, $q = 1$; and $q = p/(p-1)$ otherwise (§6.7).} Then $fg \in \mathcal{L}^1$ and $\|fg\|_1 \leq \|f\|_p \|g\|_q$; this is Hölder's inequality when $1 < p < +\infty$ (6.7.2), and is obvious in the extremal cases $p = 1$, $q = +\infty$ and $p = +\infty$, $q = 1$.

9.5.7. *Definition.* Fix a continuous function

$$K : I \times I \rightarrow \mathbb{C}.$$

We are going to use K to define a linear mapping $T : \mathcal{L}^p \rightarrow \mathcal{C}$, of which K is called the **kernel function**. Since $I \times I$ is compact for the Euclidean metric topology (see 6.1.23 or §6.1, Exercise 10) and K is continuous, K is bounded (6.8.12); as in 6.8.12, we write

$$\|K\|_{\infty} = \sup\{|K(x, y)| : x, y \in I\},$$

the sup-norm of K in $\mathcal{C}_{\mathbb{C}}(I \times I)$.

9.5.8. *Lemma.* Let K be the function of the preceding definition and let $f \in \mathcal{L}^p$.

(i) For each $x \in I$, the function

$$y \mapsto K(x, y)f(y) = K_x(y)f(y) \quad (y \in I)$$

is Lebesgue-integrable, and

$$(1) \quad \|K_x f\|_1 \leq \|K_x\|_q \|f\|_p \quad \text{for all } x \in I,$$

where q is the exponent complementary to p (as in 9.5.6, (iii)).

(ii) The function $g : I \rightarrow \mathbb{C}$ defined by

$$(2) \quad g(x) = \int K(x, y)f(y)dy \quad (x \in I)$$

satisfies the inequality

$$(3) \quad |g(x) - g(x')| \leq \|K_x - K_{x'}\|_q \|f\|_p \quad \text{for all } x, x' \in I.$$

(iii) g is continuous.

(iva) If $p > 1$ then $\|g\|_\infty \leq \|K\|_\infty (b-a)^{1/q} \|f\|_p$.

(ivb) If $p = 1, q = +\infty$ then $\|g\|_\infty \leq \|K\|_\infty \|f\|_1$.

(v) Regarding g as an element of \mathcal{L}^q , and K as an element of $\mathcal{L}^q_C(I \times I, \mathcal{S} \times \mathcal{S}, \lambda \times \lambda)$,

$$\|g\|_q \leq \|K\|_q \|f\|_p.$$

Proof. Recall that if $p = 1$ then $q = +\infty$; if $1 < p < +\infty$ then $q = p/(p-1)$; and if $p = +\infty$ then $q = 1$.

(i) Here K_x is the x -section of K (7.3.1), $K_x(y) = K(x, y)$. For each $x \in I$, $K_x \in \mathcal{C} \subset \mathcal{L}^q$, thus the inequality (1) is immediate from 9.5.6, (iii).

(ii) For all $x, x' \in I$,

$$\begin{aligned} |g(x) - g(x')| &= \left| \int K_x(y)f(y)dy - \int K_{x'}(y)f(y)dy \right| \\ &= \left| \int (K_x - K_{x'})(y)f(y)dy \right| \\ &\leq \int |(K_x - K_{x'})(y)f(y)|dy \\ &= \|(K_x - K_{x'})f\|_1 \leq \|K_x - K_{x'}\|_q \|f\|_p \end{aligned}$$

by 9.5.6, (iii), which proves (3).

(iii) Since the closed interval I is compact for the usual absolute-value metric topology (6.1.9), the product space $I \times I$ is compact for the topology generated by the max-metric d of §6.1, Exercise 10, (i) (equal, by 3.3.7, to the topology generated by the Euclidean metric). Since $I \times I$ is compact and $K : I \times I \rightarrow \mathbb{C}$ is continuous, K is uniformly continuous with respect to the indicated metrics (6.3.7). Thus, given any $\epsilon > 0$, there exists a $\delta > 0$ such that for x, y, x', y' in I ,

$$d((x, y), (x', y')) \leq \delta \Rightarrow |K(x, y) - K(x', y')| \leq \epsilon,$$

that is,

$$|x - x'| \leq \delta \ \& \ |y - y'| \leq \delta \Rightarrow |K(x, y) - K(x', y')| \leq \epsilon.$$

In particular, letting $y = y'$ we see that, for $x, x' \in I$,

$$|x - x'| \leq \delta \Rightarrow |K(x, y) - K(x', y)| \leq \epsilon \quad \text{for all } y \in I,$$

in other words,

$$(*) \quad |x - x'| \leq \delta \quad \Rightarrow \quad |(K_x - K_{x'})(y)| \leq \epsilon \quad \text{for all } y \in I.$$

If $p > 1$ (hence $1 \leq q < +\infty$), integration of the q 'th powers of the inequality on the right yields

$$\int |K_x - K_{x'}|^q d\lambda \leq \epsilon^q (b - a),$$

whence $\|K_x - K_{x'}\|_q \leq \epsilon(b - a)^{1/q}$, thus the implication $(*)$ yields

$$|x - x'| \leq \delta \quad \Rightarrow \quad \|K_x - K_{x'}\|_q \leq \epsilon(b - a)^{1/q};$$

combining this with the inequality (3), we have

$$(4) \quad |x - x'| \leq \delta \quad \Rightarrow \quad |g(x) - g(x')| \leq \epsilon(b - a)^{1/q} \|f\|_p,$$

whence the (uniform) continuity of g .

If $p = 1$, $q = +\infty$, the implication $(*)$ says that $\|K_x - K_{x'}\|_\infty \leq \epsilon$ whenever $|x - x'| \leq \delta$; combining this with the inequality (3), we have

$$(5) \quad |x - x'| \leq \delta \quad \Rightarrow \quad |g(x) - g(x')| \leq \epsilon \|f\|_1,$$

so again g is continuous. (Incidentally, (4) and (5) are in harmony under the convention $1/+\infty = 0$.)

(iva) Suppose $p > 1$. For all $x \in I$, by (1) we have

$$\begin{aligned} |g(x)| &= \left| \int K(x, y) f(y) dy \right| \\ &\leq \int |K(x, y) f(y)| dy \\ &= \|K_x f\|_1 \leq \|K_x\|_q \|f\|_p; \end{aligned}$$

but $|K_x| \leq \|K\|_\infty$ on I , whence (raise to the q 'th power, integrate, then take $(1/q)$ 'th power)

$$\|K_x\|_q \leq \|K\|_\infty (b - a)^{1/q},$$

therefore $|g(x)| \leq \|K\|_\infty (b - a)^{1/q} \|f\|_p$ for all $x \in I$, whence

$$\|g\|_\infty \leq \|K\|_\infty (b - a)^{1/q} \|f\|_p.$$

(ivb) Suppose $p = 1$ (and $q = +\infty$). For all $x \in I$, by (1) we have

$$|g(x)| \leq \|K_x f\|_1 \leq \|K_x\|_\infty \|f\|_1 \leq \|K\|_\infty \|f\|_1,$$

therefore $\|g\|_\infty \leq \|K\|_\infty \|f\|_1$.

(v) Suppose first that $p > 1$. For all $x \in I$, we have $|g(x)| \leq \|K_x\|_q \|f\|_p$, thus

$$(**) \quad |g(x)|^q \leq \left(\int |K(x, y)|^q dy \right) (\|f\|_p)^q \quad \text{for all } x \in I.$$

The function $|K|^q : I \times I \rightarrow \mathbb{C}$ is continuous, hence Borel (9.4.2, (i)), hence measurable with respect to $\mathcal{S} \times \mathcal{S}$ (9.4.2, (iv)), and it is bounded, therefore $|K|^q$ is $\lambda \times \lambda$ -integrable, so by Fubini's theorem (7.4.8) the iterated integral $\iint |K(x, y)|^q dy dx$ exists and is equal to $\int |K|^q d(\lambda \times \lambda)$; integration of (**) therefore yields

$$(\|g\|_q)^q \leq \left(\int |K|^q d(\lambda \times \lambda) \right) (\|f\|_p)^q,$$

whence the desired inequality $\|g\|_q \leq \|K\|_q \|f\|_p$.

For $p = 1$, $q = +\infty$, the same inequality holds by (ivb). \diamond

9.5.11. *Definition.* In view of (iii) of the preceding lemma, we may define a mapping $T : \mathcal{L}^p \rightarrow \mathcal{C}$ by the formula

$$(Tf)(x) = \int K(x, y)f(y)dy \quad (x \in I)$$

for all $f \in \mathcal{L}^p$. Inasmuch as $\mathcal{C} \subset \mathcal{L}^r$ for all $r \in [1, +\infty]$, the same formula may be used to define mappings $T : \mathcal{L}^p \rightarrow \mathcal{L}^r$, $T : \mathcal{C} \rightarrow \mathcal{L}^r$ and $T : \mathcal{C} \rightarrow \mathcal{C}$.

9.5.12. *Theorem.* The mapping $T : \mathcal{L}^p \rightarrow \mathcal{C}$ defined above is a compact linear mapping (where \mathcal{L}^p is equipped with the seminorm $f \mapsto \|f\|_p$, and \mathcal{C} is equipped with the sup-norm).

Proof. Linearity is obvious. Continuity will follow from compactness by (3) of 9.5.5, but here is a more direct proof: if $B = \{f \in \mathcal{L}^p : \|f\|_p \leq 1\}$ then, by (iva) and (ivb) of the lemma, the set $\{\|Tf\|_\infty : f \in B\}$ is bounded by $\|K\|_\infty(b-a)^{1/q}$ or by $\|K\|_\infty$, according as $p > 1$ or $p = 1$, thus T is continuous by (1) of 9.5.5.

To prove that T is compact, we are to show that every sequence in $T(B)$ has a uniformly convergent subsequence (9.5.5, (2)), in other words (since \mathcal{C} is complete) that $T(B)$ is a totally bounded subset of \mathcal{C} (6.1.24). By Ascoli's theorem (cf. 8.1.13), we need only show that (a) $T(B)$ is pointwise totally bounded on I , and (b) $T(B)$ is equicontinuous on I .

(a) By (iva) and (ivb) of the lemma, for each (in fact, all) $x \in I$ the set

$$\{|(Tf)(x)| : f \in \mathcal{L}^p\}$$

is a subset of the compact interval $[0, c]$, where $c = \|K\|_\infty(b-a)^{1/q}$ or $c = \|K\|_\infty$, according as $p > 1$ or $p = 1$, whence pointwise (in fact, uniform) total boundedness.

(b) Given any $\epsilon > 0$, choose $\delta > 0$ as in the proof of (iii) of the lemma. Then, by (4) and (5) of the proof, either

$$|x - x'| \leq \delta \Rightarrow |(Tf)(x) - (Tf)(x')| \leq \epsilon(b-a)^{1/q} \quad \text{for all } f \in B$$

or

$$|x - x'| \leq \delta \Rightarrow |(Tf)(x) - (Tf)(x')| \leq \epsilon \quad \text{for all } f \in B,$$

according as $p > 1$ or $p = 1$, whence the equicontinuity of $T(B)$. That the equicontinuity is uniform is not news (8.1.10). \diamond

9.5.13. Corollary. Let $1 \leq r \leq +\infty$. The linear mappings (i) $T: \mathcal{C} \rightarrow \mathcal{C}$, (ii) $T: \mathcal{C} \rightarrow \mathcal{L}^r$ and (iii) $T: \mathcal{L}^p \rightarrow \mathcal{L}^r$ defined by the formula in 9.5.11 are also compact.

Proof. If $f \in \mathcal{C}$ then $|f| \leq \|f\|_\infty \cdot 1$, therefore $\|f\|_p \leq \|f\|_\infty \|1\|_p$ (where $\|1\|_p = (b-a)^{1/p}$ when $p < +\infty$, and $\|1\|_p = 1$ when $p = +\infty$); it follows that the insertion mapping $\mathcal{C} \rightarrow \mathcal{L}^p$ is continuous. Consider the diagram

$$\mathcal{C} \xrightarrow{i_p} \mathcal{L}^p \xrightarrow{T} \mathcal{C} \xrightarrow{i_r} \mathcal{L}^r$$

where T is the linear mapping of the Theorem and i_p, i_r are the insertion mappings. The mappings contemplated in the present corollary are (i) $T \circ i_p = T|_{\mathcal{C}}$, (ii) $i_r \circ T \circ i_p$, and (iii) $i_r \circ T$; since it is clear that the composite of a continuous linear mapping and a compact linear mapping (in either order) is compact, the corollary is immediate from the Theorem. \diamond

The case $p = q = 2$ is especially transparent: $T: \mathcal{L}^2 \rightarrow \mathcal{L}^2$ leads to a compact operator \dot{T} in the Hilbert space $L^2 = L^2(I, S, \lambda)$ (see Exercise 2) with $\|\dot{T}\| \leq (\iint |K(x, y)|^2 dx dy)^{1/2}$ by (v) of 9.5.8. The theory of operators in Hilbert space is especially well-developed (cf. Exercise 5).

Exercises

1. Let $(E, \|\cdot\|)$ be a seminormed space and let $N = \{x \in E : \|x\| = 0\}$. As noted in § 6.4, Exercise 1, N is a linear subspace of E , and the mapping $x + N \mapsto \|x\|$ on the quotient vector space E/N (well-defined because $\|x + z\| = \|x\|$ for all $z \in N$) is a norm. We abbreviate $\dot{E} = E/N$ and $\dot{x} = x + N$; thus \dot{E} is a normed space with norm $\|\dot{x}\| = \|x\|$ for all $x \in E$.

(i) A sequence (x_n) in E is Cauchy in the sense of (9.5.2) if and only if (\dot{x}_n) is Cauchy in \dot{E} , and $x_n \rightarrow x$ in E if and only if $\dot{x}_n \rightarrow \dot{x}$ in \dot{E} .

(ii) E is complete in the sense of 9.5.2 if and only if \dot{E} is a Banach space.

2. With notations as in Definition 9.5.4, let $T: E \rightarrow F$ be a linear mapping, and form the quotient normed spaces \dot{E}, \dot{F} by the method of Exercise 1.

(i) If, for $x \in E$, $\|x\| = 0 \Rightarrow \|Tx\| = 0$ (cf. 9.5.5, (4)), then there exists a linear mapping $\dot{T}: \dot{E} \rightarrow \dot{F}$ such that $\dot{T}\dot{x} = (\dot{T}x)$ for all $x \in E$.

(ii) With T as in (i), T is continuous in the sense of Definition 9.5.4 if and only if \dot{T} is continuous.

(iii) With T as in (i), T is compact in the sense of Definition 9.5.4 if and only if \dot{T} is compact.

(iv) In particular, if $T: E \rightarrow F$ is any continuous (compact) linear

mapping then there exists a continuous (compact) linear mapping $\dot{T}: \dot{E} \rightarrow \dot{F}$ such that $\dot{T}\dot{x} = (\dot{T}x)^\cdot$ for all $x \in E$.

3. (i) Let (A_n) be a sequence of pairwise disjoint Lebesgue-measurable subsets of $I = [a, b]$ such that $\lambda(A_n) > 0$ for all n , and let $f_n = \varphi_{A_n}$ be the characteristic function of A_n . Then $\|f_n\|_\infty = 1$ for all n and $\|f_m - f_n\|_\infty = 1$ when $m \neq n$. Infer that the identity mapping $\mathcal{L}^\infty \rightarrow \mathcal{L}^\infty$, though continuous, is not compact.

(ii) Adapt (i) to \mathcal{L}^p ($1 \leq p < +\infty$) by suitably modifying the functions f_n .

(iii) (*Theorem of F. Riesz*) In order that the identity mapping on a normed space be compact, it is necessary and sufficient that the space be finite-dimensional.²

4. (i) When $1 \leq p < +\infty$, every $f \in \mathcal{L}^p$ is the limit in mean (of order p) of a sequence (f_n) in \mathcal{C} .³

(ii) The corresponding statement is false for $p = +\infty$, since a sequence (f_n) in \mathcal{C} with $\|f_m - f_n\|_\infty \rightarrow 0$ is uniformly convergent to a function in \mathcal{C} —and there exist functions in \mathcal{L}^∞ that are not equal a.e. to a continuous function (for example, the characteristic function of an interval $[a, c]$ with $a < c < b$).

5. With notations as in Corollary 9.5.13, let $p = q = 2$. The following conditions on the compact linear mapping $T: \mathcal{L}^2 \rightarrow \mathcal{L}^2$ are equivalent:

$$(a) \int (Tf)\bar{g}d\lambda = \int f\overline{(Tg)}d\lambda \text{ for all } f, g \in \mathcal{L}^2;$$

$$(b) \int (Tf)\bar{g}d\lambda = \int f\overline{(Tg)}d\lambda \text{ for all } f, g \in \mathcal{C};$$

$$(c) \int (Tf)\bar{g}d\lambda = \int f\overline{(Tg)}d\lambda \text{ for all polynomial functions } f, g \text{ on } I;$$

$$(d) K(x, y) = \overline{K(y, x)} \text{ for all } x, y \in I.$$

Such a mapping T is said to be **self-adjoint**, and the mapping \dot{T} it induces in the Hilbert space L^2 has an explicit representation ("Spectral Theorem") in terms of its eigenvalues.⁴

² Cf. the author, *Lectures in functional analysis and operator theory* [Springer, New York, 1974], p. 91, (23.10).

³ Cf. the author, *Measure and integration* [Macmillan, New York, 1965; reprinted Chelsea, New York, 1970], p. 220, Theorem 2.

⁴ Cf. the author, *Introduction to Hilbert space* [Oxford University Press, New York, 1961; reprinted Chelsea, New York, 1976], p. 186, Theorem 6.

Bibliography

- ABIAN, A., *The theory of sets and transfinite arithmetic*, Saunders, Philadelphia, 1965.
- ASPLUND, E. AND BUNGART, L., *A first course in integration*, Holt, Rinehart and Winston, New York, 1966.
- BARTLE, R.G., *The elements of integration*, Wiley, New York, 1966.
- BEAR, H.S., *A primer of Lebesgue integration*, Academic, New York, 1995.
- BERBERIAN, S.K., *Introduction to Hilbert space*, Oxford, New York, 1961; reprinted Chelsea, New York, 1976.
- BERBERIAN, S.K., *Measure and integration*, Macmillan, New York, 1965; reprinted Chelsea, New York, 1970.
- BERBERIAN, S.K., *Lectures on functional analysis and operator theory*, Springer-Verlag, New York, 1974.
- BERBERIAN, S.K., *A first course in real analysis*, Springer-Verlag, New York, 1994.
- BIRKHOFF, G. AND ROTA, G.C., *Ordinary differential equations*, 3rd edn., Wiley, New York, 1978.
- BOURBAKI, N., *Intégration. Ch. 5*, Hermann, Paris, 1967.
- BOURBAKI, N., *General topology. I, II*, Addison-Wesley, Reading, 1966; reprinted Springer-Verlag, New York, 1988.
- CRONIN, J., *Differential equations: Introduction and qualitative theory*, 2nd edn., Marcel Dekker, New York, 1994.
- DEDEKIND, R., *Essays on the theory of numbers* (translated from the German original), Open Court, LaSalle, 1901; reprinted Dover, New York.
- DIXMIER, J., *C*-algebras*, North-Holland, Amsterdam, 1977.
- DIXMIER, J., *General topology*, Springer-Verlag, New York, 1984.
- GILLMAN, L. AND JERISON, M., *Rings of continuous functions*, Van Nostrand, Princeton, 1960; reprinted Springer-Verlag, New York, 1976.
- HALMOS, P.R., *Measure theory*, Van Nostrand, New York, 1950; reprinted Springer-Verlag, New York, 1974.
- HALMOS, P.R., *Naive set theory*, Van Nostrand, Princeton, 1960; reprinted Springer-Verlag, New York, 1974.

- HARTMAN, P., *Ordinary differential equations*, 2nd edn., Birkhäuser, Boston, 1982.
- HAUSDORFF, F., *Set theory*, 3rd edn., Chelsea, New York, 1957.
- HEWITT, E. AND STROMBERG, K., *Real and abstract analysis*, Springer-Verlag, New York, 1965.
- HILDEBRANDT, T.H., *Introduction to the theory of integration*, Academic, New York, 1963.
- HOBSON, E.W., *The theory of functions of a real variable and the theory of Fourier's series. Vol. 1*, Dover, New York, 1957.
- KADISON, R.V. AND RINGROSE, J.R., *Fundamentals of the theory of operator algebras. Vols. I-IV*, Academic, New York, 1983–1992.
- KAPLANSKY, I., *Set theory and metric spaces*, 2nd edn., Chelsea, New York, 1977.
- KESTELMAN, H., *Modern theories of integration*, Oxford, 1937; 2nd revised edn., Dover, New York, 1960.
- KURATOWSKI, K., *Topologie. I.*, Monografie Matematyczne, 2nd edn., Warsaw, 1948.
- LANDAU, E., *Foundations of analysis*, Chelsea, New York, 1951.
- LANDAU, E., *Differential and integral calculus*, Chelsea, New York, 1951.
- LOOMIS, L.H., *An introduction to abstract harmonic analysis*, Van Nostrand, New York, 1953.
- MC SHANE, E.J., *Integration*, Princeton, 1944.
- OXToby, J., *Category and measure*, Springer-Verlag, New York, 1971.
- RICKART, C.E., *General theory of Banach algebras*, Van Nostrand, Princeton, 1960; reprinted R.E. Krieger, Huntington, 1974.
- ROYDEN, H., *Real analysis*, 3rd edn., Macmillan, New York, 1988.
- RUDIN, W., *Principles of mathematical analysis*, 3rd edn., McGraw-Hill, New York, 1976.
- SUPPES, P., *Axiomatic set theory*, Van Nostrand, Princeton, 1960; reprinted Dover, 1972.
- SZ.-NAGY, B., *Introduction to real functions and orthogonal expansions*, Oxford, New York, 1965.

Index of Notations

SYMBOL	PAGE	SYMBOL	PAGE
$x \in A$	2	$\limsup a_n$	79
$\mathbb{P}, \mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$	2	$\liminf a_n$	80
$A \subset B$	3	$\lim_{n \rightarrow \infty} a_n$	82
$A \cup B, A \cap B$	4	$\lambda^*(A)$	87
$A', A - B$	4	$\mathcal{M}(\lambda^*)$	96
$\{a\}$	8	$\lambda_i(A), \lambda_e(A)$	97
$f : X \rightarrow Y$	9	Γ	99
φ_A	10	$\mathcal{S}(\mathcal{E}), \mathcal{A}(\mathcal{E})$	101
$g \circ f$	10	(X, \mathcal{S}, μ)	103
$\mathcal{P}(X)$	12	$A \Delta B$	107
$\bigcup_{i \in I} A_i, \bigcap_{i \in I} A_i$	13	$\mathcal{B}_{\mathbb{R}}(T), \mathcal{B}_{\mathbb{C}}(T)$	119
$\prod_{i \in I} X_i$	13	$\ x\ _{\infty}, d_{\infty}(x, y)$	119
pr_j	14	$\ x\ _p$	119
X/\sim	19	$d_p(x, y)$	122
$X \approx Y$	23	$\lim_{n \rightarrow \infty} x_n$	124
$E \sim F$	46	$x_n \rightarrow x$	124
$E \lesssim F$	47	$B_r(c)$	126
$\text{card } E$	49	$U_r(c)$	127
\aleph_0, c	49	$S_r(c)$	129
$\text{card } E \leq \text{card } F$	50	\overline{A}	134
$\mathcal{F}(E, F), F^E$	56	A°	137
$\text{ord } E$	59	$\text{int } A, \text{ext } A$	137
ω	60	$\partial A, \text{bd } A$	137
$\text{ord } E \leq \text{ord } F$	62	$\lim_{x \rightarrow c, x \in A} f(x)$	142
\aleph_1, Ω	70	$\lim_{x \rightarrow c} f(x)$	142
$\overline{\mathbb{R}}$	73	$f(c+), f(c-)$	144
$\sum_{i \in I} a_i$	74	$f'(c), f'_l(c), f'_r(c)$	145
$\sum_{k=1}^{\infty} a_k$	76	$\mathcal{B}(X)$	149
$x_i \uparrow x, x_i \downarrow x$	78	$f \cup g, f \cap g$	151

SYMBOL	PAGE	SYMBOL	PAGE
$f^+, f^-, f $	152	$\text{diam } A$	278
$\sup_n f_n, \inf_n f_n$	152	$\mathcal{B}(T, Y)$	289
$\limsup f_n$	153	$D(f, g)$	290
$\liminf f_n$	153	$\mathcal{C}(X, Y)$	292
$N(f), \text{ISF}$	160	$\mathcal{C}_{\mathbb{R}}(X), \mathcal{C}_{\mathbb{C}}(X)$	292
$\mathcal{L}_{\mathbb{R}}^1(\mu)$	167	$\mathcal{L}_{\mathbb{C}}^1$	312
$\int f d\mu$	167	$\ f\ _1$	314
$\int_E f d\mu$	172	$\mathcal{L}_{\mathbb{C}}^1$	317
$\int_a^b f d\lambda$	172	$\mathcal{L}_{\mathbb{C}}^{\infty}, \ f\ _{\infty}$	323
$f \cdot \mu$	184	$L_{\mathbb{C}}^{\infty}$	324
$\nu \ll \mu$	192	$\mathcal{L}_{\mathbb{C}}^p, \ f\ _p$	332
$V_a^b f$	202	$L_{\mathbb{C}}^p$	336
$\sigma \prec \tau$	202	$\text{Re } f, \text{Im } f$	349
$\langle I \rangle, f \langle I \rangle $	205	\mathbb{U}	362
$\mathcal{BV}[a, b]$	212	$\mathcal{C}_0(X)$	363
$\mathcal{AC}[a, b]$	213	$\mu^*(A)$	365
$\limsup_{x \rightarrow c, x \in A} f(x)$	216	$\mathcal{S} \times \mathcal{T}$	371
$\liminf_{x \rightarrow c, x \in A} f(x)$	216	$\mu \times \nu$	378
D^+, D_+, D^-, D_-	220	$\iint h d\nu d\mu$	384
$\limsup_{x \rightarrow c} f(x)$	233	μ_A	389
$\liminf_{x \rightarrow c} f(x)$	233	$ \nu $	431
$\underline{D}, \overline{D}$	236	$\mu \perp \nu$	444
f_*, f^*	267	$f * g$	458

Index

- Absolute continuity of measures, 192, 322, 440
 - of signed measures, 440, 445
- Absolutely continuous function, 206, 213
- Adherence of a subset of a topological space, 134
- Adherent point, 133
- a.e., 157
- a.e. primitive, 248
- a.e. unique, 157
- algebra of sets, 101
- Algebra, 212
- Almost everywhere, 157
- Arzela–Young theorem, 113
- Ascoli’s theorem, 401, 407
- Axiom of Choice, 13

- Baire category theorem, 304
- Baire space, 304
- Banach algebra, 331, 352
- Banach space, 318, 461
- Base for a topology, 279
- Basis of a vector space, 43
- Bijection, 10
- Bijective function, 10
- Bilateral sequence, 11
- Bilinear mapping, 344
- Boolean ring, 432
 - of sets, 107
- Borel function, 151, 453
- Borel sets of \mathbb{R} , 102
 - of a topological space, 149
- Borel space, 149
- Boundary of a subset of a topological space, 137
- Boundary point, 137

- Bounded above, 22
- Bounded below, 22
- Bounded function, 119, 288
- Bounded set, in a metric space, 298
 - in a pre-ordered set, 22
- Bounded variation, function of, 202, 264

- C^* -algebra, 352
- Canonical factorization of a function, 20
- Cantor set, 99
- Cantor’s theorem, 50
- Cardinal arithmetic, 52
- Cardinal numbers, 49
 - countable, 52
 - finite, 52
 - infinite, 52
 - uncountable, 52
- Cardinality, 49
 - finite, 37
 - of an ordinal number, 68
- Cartesian product, 5
- Cauchy sequence, 278
- Cauchy’s criterion for convergence, 31
 - for uniform convergence, 287, 288
- Cavalieri’s principle, 397
- Chain in an ordered set, 21
- Characteristic function of a subset, 10
- Choice function, 14
- Closed ball in a metric space, 126
- Closed set in a metric space, 126
 - in a topological space, 131
- Closure of a subset of a topological space, 134
- Commutative diagram, 20
- Compact linear mapping, 461

- Compact space, 275
 - subset, 276
- Complement of a subset, 4
- Complete measure space, 108, 156
 - metric space, 282
 - ordered field, 26, 33
 - seminormed space, 461
- Completion of a metric space, 309
- Complex measure, 319
 - numbers, 2, 118, 350
- Composite function, 10
- Composition of functions, 10
 - of relations, 8
- Conclusion of a proposition, 3
- Congruence of integers, 17
- Continuity at a point, 139
- Continuous function, 139
 - linear mapping, 461
- Continuum Hypothesis, 58
- Continuum, cardinal of, 49, 57
- Contraction of a measure to a subset, 389
- Contradiction, law of, 5
- Contrapositive form of a theorem, 3
- Convergent sequence in $\overline{\mathbb{R}}$, 82
 - in a metric space, 124
 - sequence in a topological space, 137
- Convergent sequence of functions,
 - pointwise, 153, 286
- Converse of a proposition, 3
- Convolution of functions, 458
- Coordinate, 13
- Countable set, 38
- Countably infinite set, 39
- Covering, 274

- De Morgan's formulas, 13
- Decreasing net, 77
- Dedekind cut, 32
- Dense subset of a topological space, 136
- Denumerable set, 39
- Denumerably infinite set, 39
- Derivates of a function, 220, 236
- Derivative of a function, 145
- Diagonal, 8, 17
- Diameter of a subset, 130, 278, 298, 299
- Difference of two subsets, 4
- Differentiable function, 145
- Differentiation, term-by-term, 296
- Dini derivates of a function, 220
- Dini's theorem, 293
- Dirac measure, 106
- Discrete measure, 106
 - metric, 118
 - metric space, 118
 - topology on a set, 131
- Distance function, 118
- Domain of a relation, 8
- Dominated convergence theorem, 176
- Dual of an order relation, 21
- Dual space, 343

- Empty function, 17
- Equi-uniformly continuous set of functions, 400
- Equicontinuous set of functions, 400
 - at a point, 399
- Equipollent sets, 46
- Equivalence class, 18
- Equivalence relation, 9, 17
 - Deduced from a function, 17
- Equivalent metrics, 131
 - propositions, 3
 - signed measures, 448
- Essential bound, 323
- Essential supremum of a function, 157, 323
- Essentially bounded function, 157, 323
- Essentially unique, 157
- Euclidean metric, 122
 - space, 122
- Excluded middle, law of, 5
- Extended real numbers, 73
 - algebraic operations on, 74
- Exterior measure, 97
- Exterior of a subset of a topological space, 137

- Family indexed by a set, 11
- Fatou's lemma, 177
- Final set of a function, 9
- Finite character, a set of sets of, 43
- Finite covering, 274
- Finite diameter, subset of, 130, 278

- Finite intersection property, 275
- Finite set, 34, 38
- Finite signed measure, 190, 319
- Finite-dimensional normed spaces,
 - F. Riesz's characterization, 468
- Finitely additive set function, 112
- First category, subset of the, 303
- First countable topological space, 133
- Frequently, 80
- Fubini's theorem, 387, 395
- Fubini-Tonelli theorem, 385, 392
- Function, 9
- Function of bounded variation,
 - 202, 264
- Fundamental sequence of
 - neighborhoods, 133
- Fundamental theorem of calculus,
 - Lebesgue's, 252
- Generated algebra, 101
 - σ -algebra, 101
- Graph of a function, 9
- Hahn decomposition of a finite signed measure, 192
- Hahn-Kolmogorov extension theorem, 367
- Hausdorff space, 146, 275
- Hausdorff's maximality principle,
 - after 42
- Heine-Borel theorem, 273
- Hilbert space, 345
- Hölder's inequality, 120, 332
- Homeomorphism, 140
- Hypothesis of a proposition, 3
- Ideal, order, 63
- Identity indexing, 11
 - mapping, 10
 - relation, 8
- Image, direct, 7
- Imaginary part of a complex-valued function, 349
- Increasing net, 77
- Indefinite integral, 184, 207, 213, 248
- Index set, 11
- Induction, principle of mathematical, 23
- Inductive partially ordered set, 42
- Infimum, 73
- Infinite set, 34, 38
- Initial ordinal of a cardinal number, 69
- Initial segment of a well-ordered set, 60
- Initial set of a function, 9
- Injection, 10
- Injective function, 10
- Inner product, 344
- Inner regularity of Lebesgue measure, 105
- Insertion mapping, 10
- Integers, 2
- Integrable function, 167, 178, 312
 - simple function, 160
- Integral, 167, 177, 312
- Interior measure, 97
- Interior of a subset of a topological space, 137
- Interior point, 127, 132
- Intersection of subsets, 4, 13
- Inverse function, 16
 - image, 7, 11
- Isometric mapping, 307
- Isometry, 307
- Iterated integral, 384
- Iterated limits theorem, 294
- Jordan decomposition of a function of
 - bounded variation, 205
- Jordan-Hahn decomposition of a signed measure, 431
- Kernel function, 463
- l.s.c., 229, 232
- \mathcal{L}^1 -norm, 314
- Lebesgue decomposition of a function
 - of bounded variation, 264
 - of signed measures, 446, 450
- Lebesgue measure, 96
- Lebesgue number of a covering, 401
- Lebesgue outer measure, 87
- Lebesgue's criterion for Riemann integrability, 268
 - Fundamental theorem of calculus, 252
 - singular function, 210

- Lebesgue-integrable function, 167
- Lebesgue-measurable function, 150
 - set, 93
- Left limit of a function, 144
- Left-derivative of a function, 145
- Left-differentiable function, 145
- Lemma on monotone classes, 180
- Lexicographic ordering of a product of
 - ordered sets, 22, 25, 45
 - reverse of, 71
- Limit inferior of a function, 216, 233
 - of a sequence, 80
- Limit of a function at a point, 142
- Limit of a sequence in $\overline{\mathbb{R}}$, 82
 - in a metric space, 124
- Limit of a sequence of functions,
 - pointwise, 153, 286
- Limit ordinal, 71
- Limit superior of a function, 216, 233
 - of a sequence, 79
- Lindelöf's theorem, 280
- Linear mapping, compact, 461
 - continuous, 461
- Linear ordering, 21
- Lipschitz condition, 206, 270, 328, 410
- Locally compact space, 363
- Lower bound, 22
- Lower derivate of a function, 236
- Lower semicontinuous function, 229, 232
- Lower semicontinuous regularization
 - of a function, 268
- \mathcal{L}^p -norm, 332
- Majorant, 22
- Mapping, 9
- Mathematical induction, principle of, 23
- Maximal element in a pre-ordered set, 22
- Meager subset, 303
- Mean, convergence in, 315
 - of order p , 461
- Measurable function, 149, 311
 - rectangle, 372
 - space, 149
 - subset, 93, 95, 149, 366
- Measure on a σ -algebra, 103
 - on an algebra of sets, 365
- Measure space, 103
- Metric, 117
- Metric space, 118
- Metric subspace, 138
- Metrizable topological space, 131
- Minimal element in a pre-ordered set, 23
- Minkowski p -metric, 122
 - p -norm, 119
 - space, 122
- Minkowski's inequality, 120, 333
- Minorant, 22
- Monotone class of sets, 179
 - lemma on, 180
- Monotone convergence theorem, 176
 - set function, 112
- Monotonic functions, 222, 253
- Mutually singular signed measures, 444
- Negation of a proposition, 2
- Negative part of a function, 152
- Negligent function, 213
- Negligible set, 87, 95, 156
- Neighborhood of a point, 128, 132
- Nested closed sets property, 299
- Nested intervals, theorem on, 32
- Net, 77
- Nonmeager subset, 304
- Nonmeasurable set, Vitali's example, 110
- Nonnegative integers, 2
- Nonoverlapping intervals, 206
- Nonzero set of a function, 160
- Norm, 314
- Normed space, 314
 - finite-dimensional, 468
- Null functions on a measure space, 314
- Null set, 156
- One-one correspondence, 10
 - function, 10
- Onto function, 10
- Open ball in a metric space, 127
- Open covering, 274
- Open set in a metric space, 128
 - in a topological space, 131
- Operator, 9

- Order ideal, 63
 - isomorphism, 23
 - monomorphism, 23
 - morphism, 23
 - relation, 20
- Ordinal numbers, 60
 - countable, 67
 - denumerable, 67
 - finite, 67
 - infinite, 67
 - natural ordering of, 62
 - sum of, 68
 - uncountable, 67
- Ordinality, 59
 - of a cardinal number, 69
- Outer measure, 95, 365
 - Lebesgue, 87
- Outer regularity of Lebesgue measure, 105
- p -norm, 332
- Partial ordering, 20
- Partially ordered set, 21
- Partition of a set, 18
- Peano's existence theorem, 416
- Picard's existence theorem, 410
- Planar Lebesgue measure, 379
- Point mass, 106
- Pointwise Cauchy sequence of
 - functions, 286
 - convergent sequence of functions. 153, 286
 - limit of a sequence of functions, 153, 286
 - totally bounded set of functions, 399
- Polar decomposition of a signed measure, 444
- Positive integers, 2
- Positive linear form, 162, 168
- Positive part of a function, 152
- Power series, 84, 288, 297, 353
- Power set, 12
- Pre-ordered set, 21
- Pre-ordering relation, 21
- Primitive, a.e., 248
- Principle of mathematical induction, 23
 - of transfinite induction, 42
- Product measure space, 378
- Product of sets, 13
 - real or complex measures, 379
 - σ -finite measures, 378
 - order relations, 22, 24
 - σ -algebras, 371
 - topological spaces, 352
- Projection mapping, 14
- Proper subset, 4
- Proposition, 2
- Pseudometric, 118, 123
 - space, 118, 123
- p th-power integrable function, 332
- Purely negative subset, 428
- Purely positive subset, 428
- Quasicompact space, 274
 - subset, 274
- Quotient mapping, 19
- Quotient of a pre-order relation, 22
- Quotient set, 19
- Radon-Nikodym theorem, 195, 441
- Range of a relation, 8
- Rare subset, 303
- Rational numbers, 2
- Real measure, 319
- Real numbers, 2
 - extended, 73
- Real part of a complex-valued function, 349
- Real variable, function of, 33
- Real-valued function, 33
- Reflexive relation, 17, 20
- Regularization of a function, lower semicontinuous, 268
 - upper semicontinuous, 268
- Relation, 6
- Relative topology induced on a subset, 138, 146, 276
- Restriction of a function, 11
- Reverse of a relation, 8
 - of an order relation, 21
- Riemann integrability, Lebesgue's criterion, 268
- Riesz representation theorem, 327, 341
- Riesz-Fischer theorem, 341
- Right limit of a function, 144

- Right-derivative of a function, 145
- Right-differentiable function, 145
- Ring of subsets, 432
- Rising sun lemma, 244
- Russell's paradox, 5

- Schröder-Bernstein theorem, 47
- Second category, subset of the, 304
- Sections of a function on a product space, 383
 - of a subset of a product space, 372
- Self-adjoint linear mapping, 468
- Semicontinuous approximations of a Lebesgue-integrable function, 239
- Semicontinuous regularization of a function, lower, 268
 - upper, 268
- Seminorm, 314
- Seminormed space, 460
 - complete, 461
- Separable metric space, 279
- Separated space, 146, 275
- Sequence, 11
- Sesquilinear form, 344
- σ -algebra of subsets, 101
- σ -finite measure on an algebra of sets, 369
- σ -ring of subsets, 432
- Signed measure, 424
 - finite, 190
- Similar pre-ordered sets, 23
- Simple function, 154
- Simple ordering, 21
- Simply ordered set, 21
- Singleton, 8
- Singular function, 264
 - Lebesgue's, 210
- Sphere in a metric space, 129
- Stone-Weierstrass theorem, 359, 361
- Subcovering, 274
- Subset, 3
- Subtractible functions, 423
 - measures, 423
- Subtractive set function, 320
- Sup-metric, 119, 290
- Sup-norm, 119
- Superset, 3
- Supremum, 73

- Surjection, 10
- Surjective function, 10
- Symmetric difference of sets, 107
- Symmetric relation, 17

- Term-by-term differentiation, 296
- Theorem on nested intervals, 32
- Theorem, 3
- Topological space, 131
- Topology, 131
 - of uniform convergence, 399
- Total ordering, 21
- Total variation of a finite signed measure, 198
 - of a function, 202
 - of a signed measure, 431
- Totally bounded metric space, 278
- Totally bounded set of functions, 399
- Transfinite induction, principle of, 42
- Transformation, 9
- Transitive relation, 17, 20
- Trichotomy, law of, 21
 - for cardinal numbers, 51
 - for ordinal numbers, 63
- Trigonometric polynomials, 363
- Trivial measure space, 106
 - relation, 18
 - topology on a set, 131
- Tukey's lemma, 43

- u.s.c., 231, 232
- Ultimately, 80
- Uncountable set, 38
- Uniform boundedness principle, 305
- Uniform convergence, 286
 - topology of, 399
- Uniform limit, 286
- Uniformly Cauchy, 286
- Uniformly continuous function, 300
- Union of subsets, 4, 13
- Unique extension theorem, 370
- Unitary space, 122
- Upper bound, 22
- Upper derivate of a function, 236
- Upper envelope of a family of functions, 230
- Upper semicontinuous function, 231, 232

- Upper semicontinuous regularization
 - of a function, 268
- Upward directed partially ordered set,
77
- Weierstrass M -test, 288
- Weierstrass approximation
theorem, 361
- Weierstrass–Bolzano property, 277
- Weierstrass–Bolzano theorem, 276
- Well-ordered set, 41, 59
- Well-ordering of a set, 41
- Well-ordering theorem, 42
- Zermelo’s theorem, 42
- Zorn’s lemma, 42

Rotman: Galois Theory
Rubel/Colliander: Entire and Meromorphic Functions
Sagan: Space-Filling Curves
Samelson: Notes on Lie Algebras
Schiff: Normal Families
Shapiro: Composition Operators and Classical Function Theory
Simonnet: Measures and Probability
Smith: Power Series From a Computational Point of View
Smoryński: Self-Reference and Modal Logic
Stillwell: Geometry of Surfaces
Stroock: An Introduction to the Theory of Large Deviations
Sunder: An Invitation to von Neumann Algebras
Tondeur: Foliations on Riemannian Manifolds
Wong: Weyl Transforms
Zong: Strange Phenomena in Convex and Discrete Geometry

Universitext



Integration theory and general topology form the basis for a first-year graduate course in real analysis. The material in the first chapter (construction of the real numbers, Zorn's Lemma, and transfinite induction) and topology are introduced and developed in increasing depth.

The treatment of integration theory is quite thorough, covering convergence theorems, product measure, the Radon–Nikodym theorem, and Lebesgue's theorem on the existence of primitive functions, while topology plays a supporting role. In the later chapters, integration is applied to topics such as function spaces, the Riesz representation theorem, and the existence theorems for an ordinary differential equation. The study of function spaces lays a firm foundation for functional analysis.

~~GST 11 1998~~

DEC 29 1999

ISBN 0-387-98480-1
www.springer-ny.com

